

Introduction to Probabilistic Machine Learning

Neill DF Campbell

Cambridge ELLIS Summer School, 13 July 2026

Department of Computer Science, University College London
(Slide input credits: Simon Prince, Mike Tipping, David MacKay)



What do you hope to take away from
this session?

Overview

Common Questions?

What questions do we have about ML?

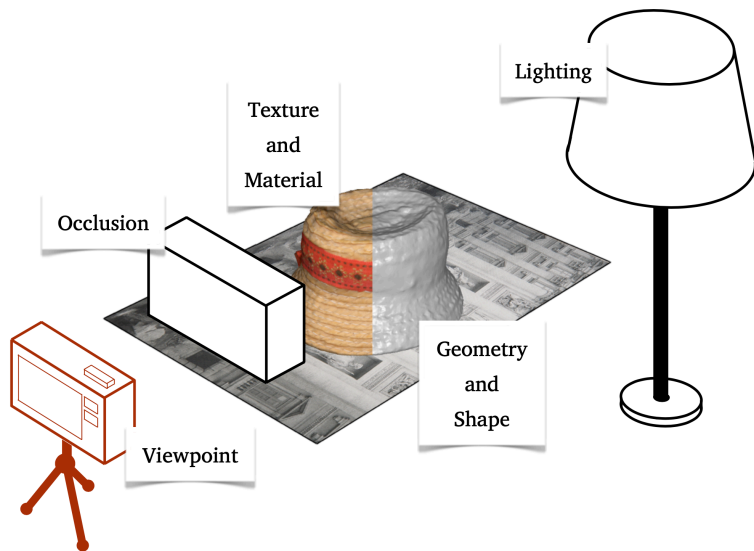
- Can I use ML to solve x ?
- What does ML actually do?
- Isn't ML just the same as y ?
- Can I replace myself/my research team with ML?
- How much data do I need?
- Can I just use Deep Learning/Generative AI/ChatGPT?
- Surely Deep Learning/Generative AI/ChatGPT is all hype?
- Can any of this be used for science/engineering?

Things to consider..

- Are all datasets equal or how to choose your data?
- Gotchas: What Machine Learning can and can't do for you.
- Average vs worst case Machine Learning
- Machine Learning and Causality
- Trade-offs in Machine Learning

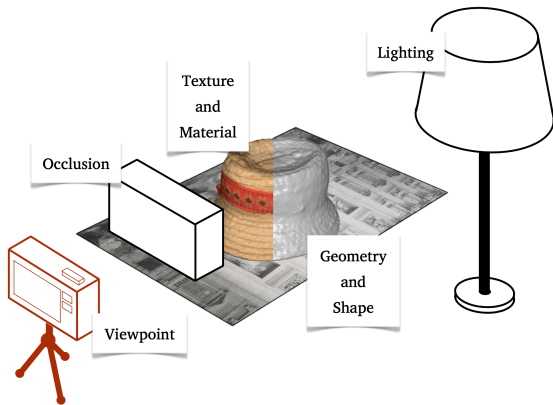
Motivation

Why do we need to care about Uncertainty in ML?



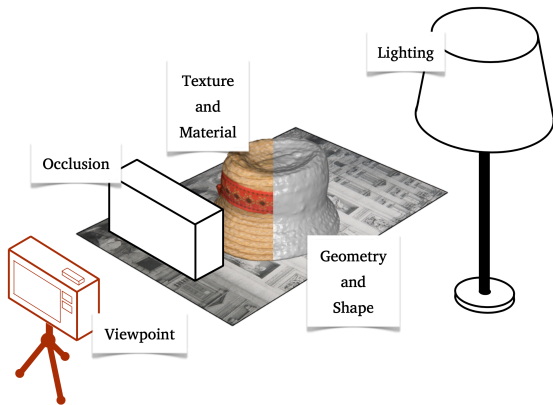
Why do we need to care about Uncertainty in ML?

- Ambiguity in the task



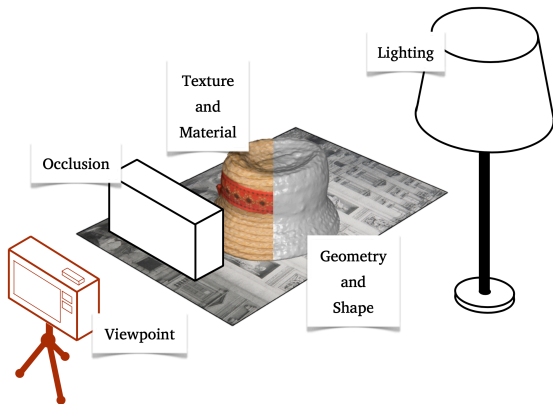
Why do we need to care about Uncertainty in ML?

- Ambiguity in the task
- Ambiguity in our models



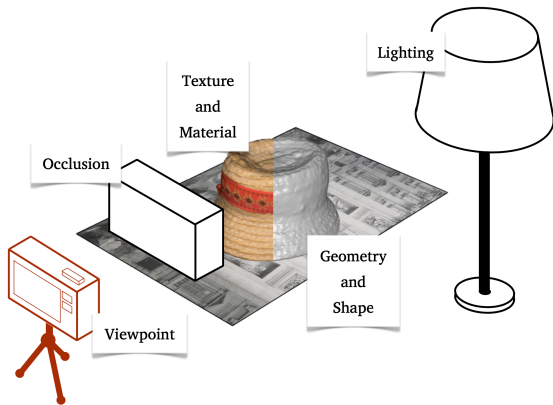
Why do we need to care about Uncertainty in ML?

- Ambiguity in the task
- Ambiguity in our models
- Downstream decision making



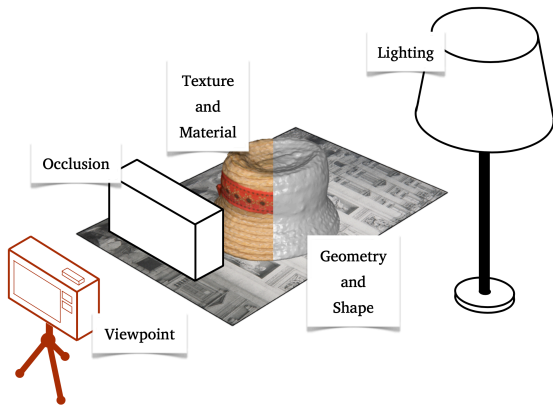
Why do we need to care about Uncertainty in ML?

- Ambiguity in the task
- Ambiguity in our models
- Downstream decision making
 - Safe



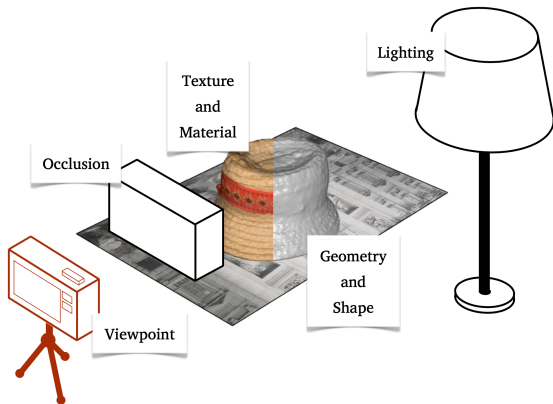
Why do we need to care about Uncertainty in ML?

- Ambiguity in the task
- Ambiguity in our models
- Downstream decision making
 - Safe
 - Robust



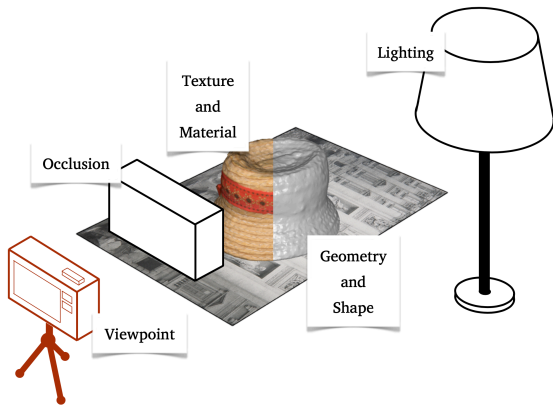
Why do we need to care about Uncertainty in ML?

- Ambiguity in the task
- Ambiguity in our models
- Downstream decision making
 - Safe
 - Robust
 - Transparent



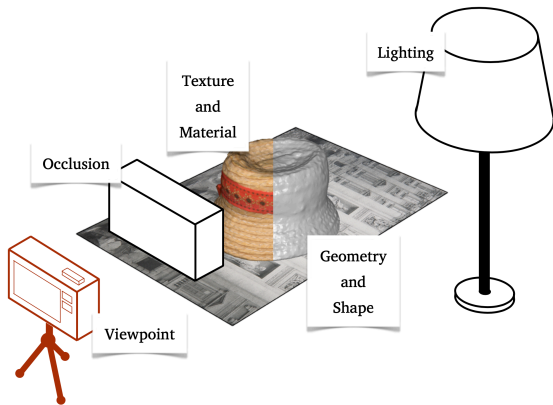
Why do we need to care about Uncertainty in ML?

- Ambiguity in the task
- Ambiguity in our models
- Downstream decision making
 - Safe
 - Robust
 - Transparent
- Improved performance



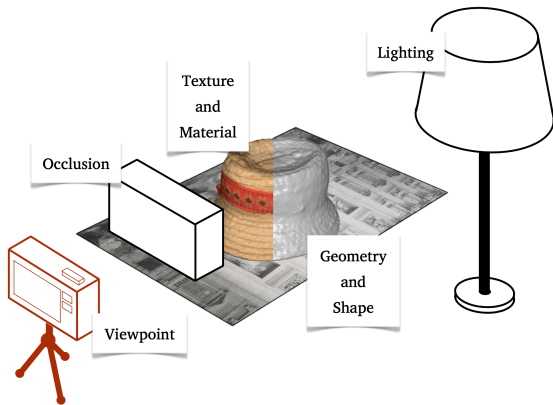
Why do we need to care about Uncertainty in ML?

- Ambiguity in the task
- Ambiguity in our models
- Downstream decision making
 - Safe
 - Robust
 - Transparent
- Improved performance
 - Data efficiency



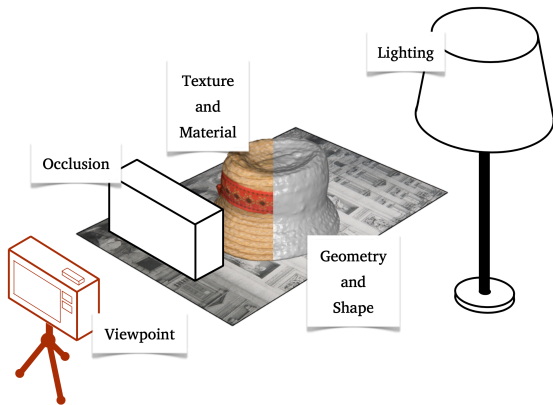
Why do we need to care about Uncertainty in ML?

- Ambiguity in the task
- Ambiguity in our models
- Downstream decision making
 - Safe
 - Robust
 - Transparent
- Improved performance
 - Data efficiency
 - Self-supervision

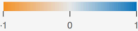
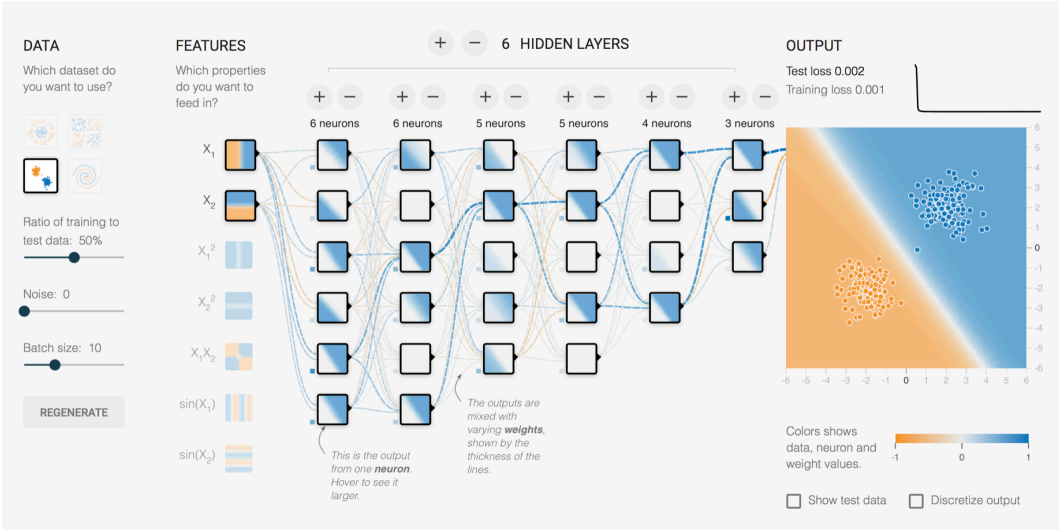


Why do we need to care about Uncertainty in ML?

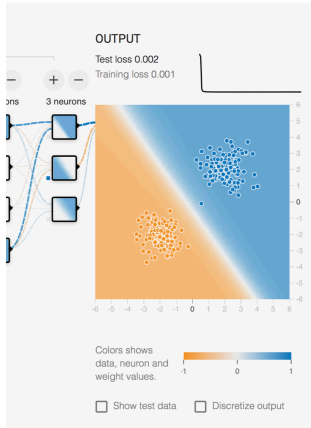
- Ambiguity in the task
- Ambiguity in our models
- Downstream decision making
 - Safe
 - Robust
 - Transparent
- Improved performance
 - Data efficiency
 - Self-supervision
- **Evaluation and Model Selection!!**



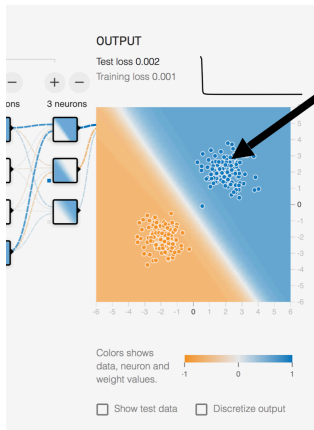
Dangers of avoiding probabilistic approaches...



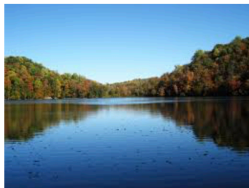
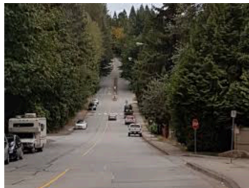
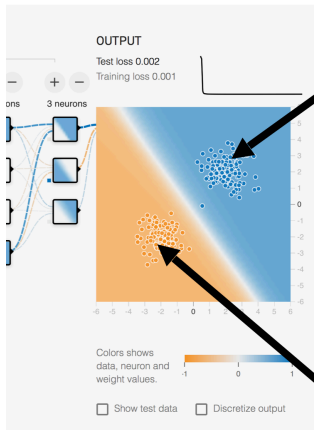
Dangers of avoiding probabilistic approaches...



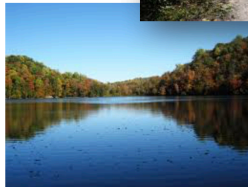
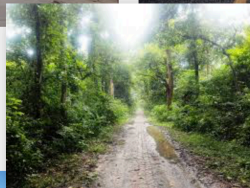
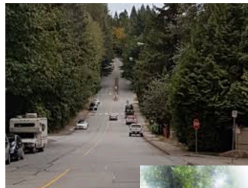
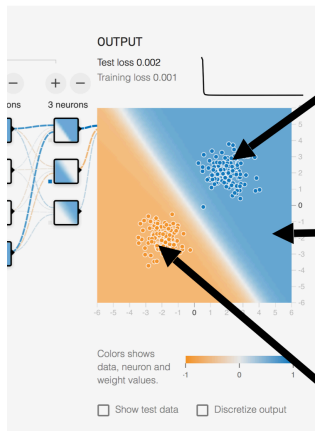
Dangers of avoiding probabilistic approaches...



Dangers of avoiding probabilistic approaches...



Dangers of avoiding probabilistic approaches...

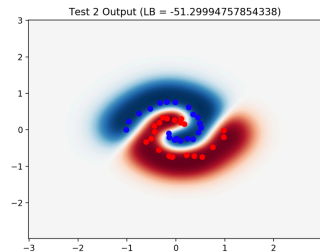
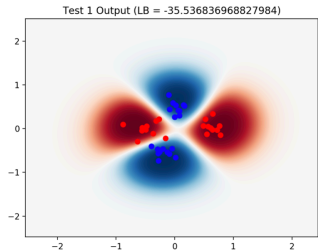
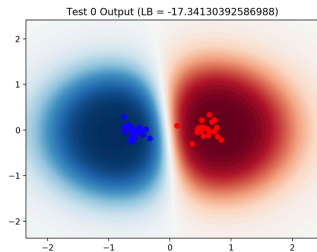


Dangers of avoiding probabilistic approaches...

What if we use a probabilistic approach?

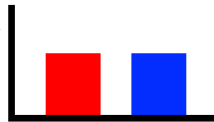
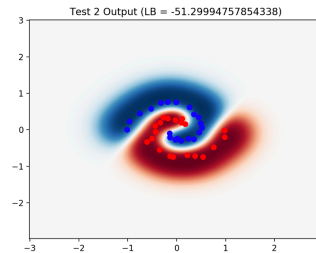
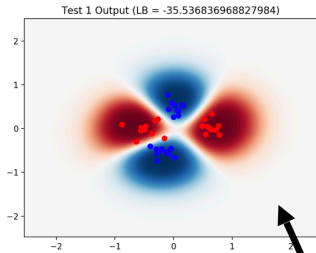
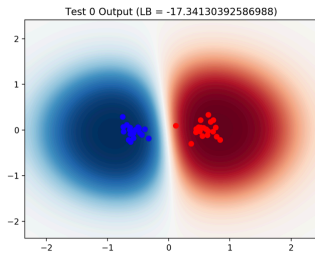
Dangers of avoiding probabilistic approaches...

What if we use a probabilistic approach?



Dangers of avoiding probabilistic approaches...

What if we use a probabilistic approach?

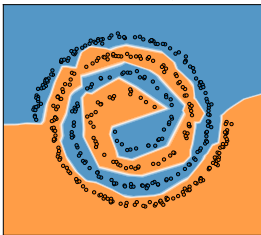


We need to consider the **properties** of
our Machine Learning
models/approaches..

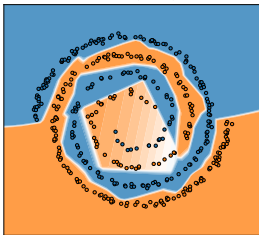
Spiral Dataset with an MLP.

Same data, varying MLP init, MLP: 32 x 2

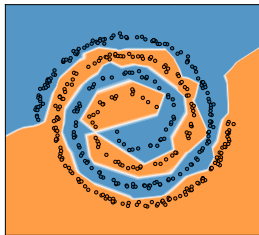
init_seed=0 (acc=1.00)



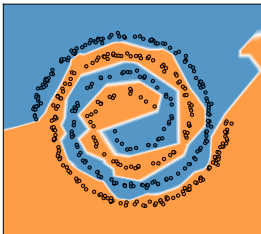
init_seed=1 (acc=0.96)



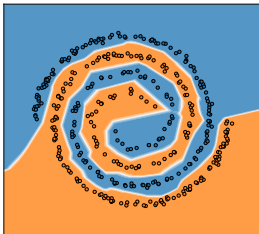
init_seed=2 (acc=1.00)



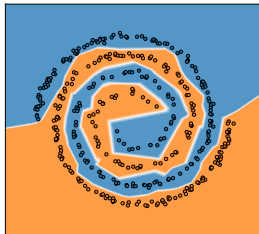
init_seed=3 (acc=1.00)



init_seed=4 (acc=1.00)



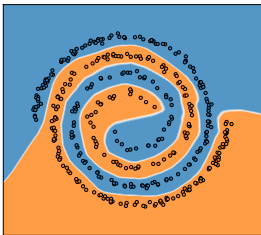
init_seed=5 (acc=1.00)



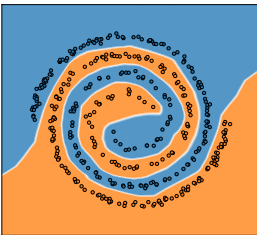
Spiral Dataset with an MLP.

Same data, varying MLP init, MLP: 64 x 4

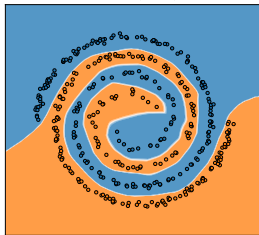
init_seed=0 (acc=1.00)



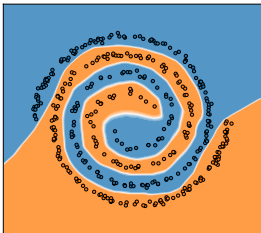
init_seed=1 (acc=1.00)



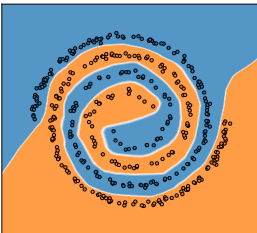
init_seed=2 (acc=1.00)



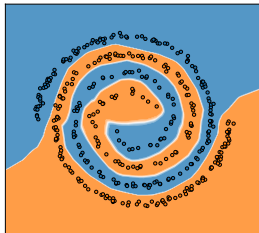
init_seed=3 (acc=1.00)



init_seed=4 (acc=1.00)



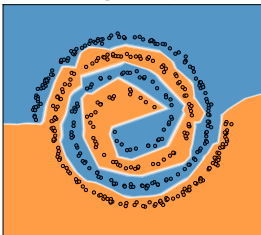
init_seed=5 (acc=1.00)



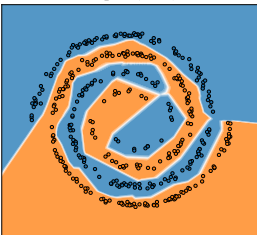
Spiral Dataset with an MLP: Different datasets..

Varying dataset, fixed MLP initialization

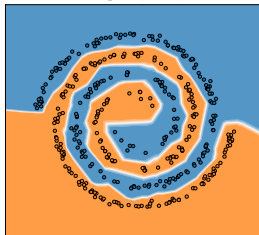
data_seed=0 (acc=1.00)



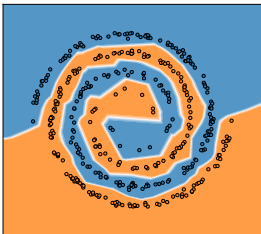
data_seed=1 (acc=1.00)



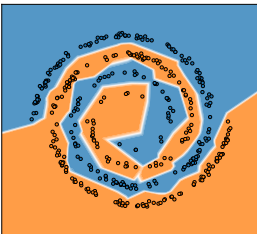
data_seed=2 (acc=1.00)



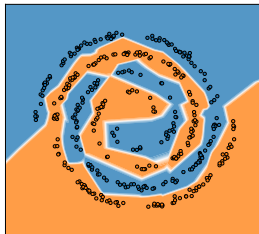
data_seed=3 (acc=1.00)



data_seed=4 (acc=1.00)

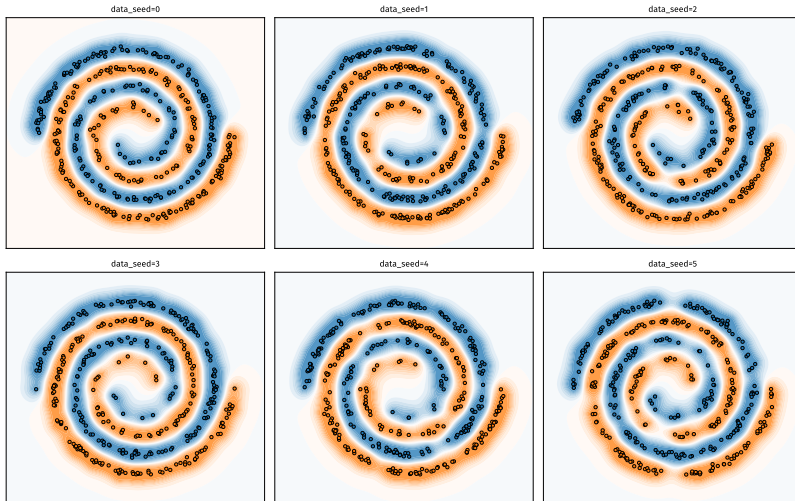


data_seed=5 (acc=1.00)



Spiral Dataset with a Bayesian (GP) model: Different datasets..

Varying dataset, GP classifier (RBF kernel, Bernoulli likelihood)



No Free Lunch

Overview...

Overview

Motivation

No Free Lunch

Whirlwind Introduction to Inverse Probabilities

Model Selection

Evaluation

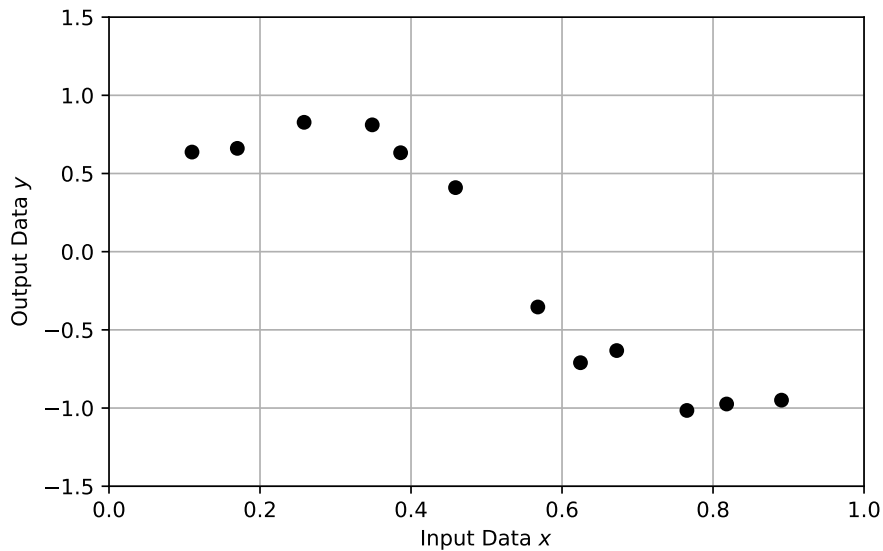
Bayesian Machine Learning: Simple Example

Why don't we do Model Selection in Machine Learning?

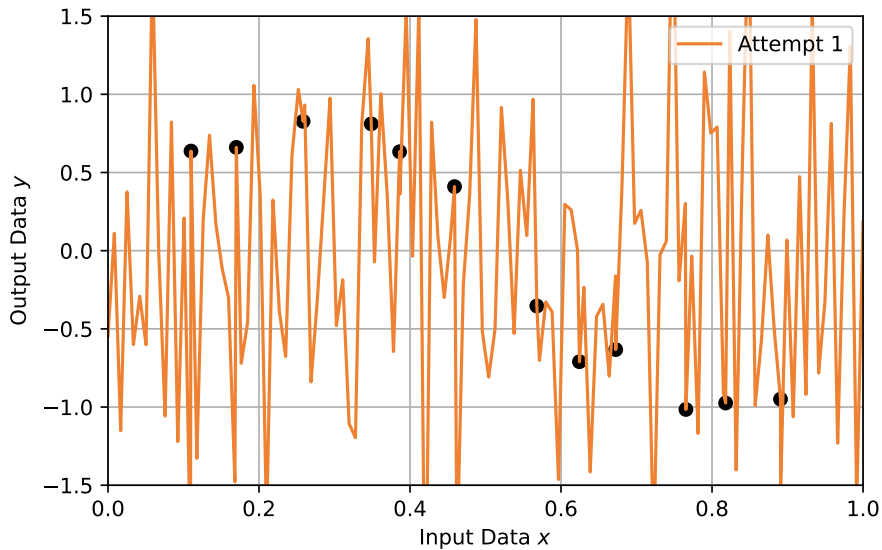
Conclusions

What happens between the dots?

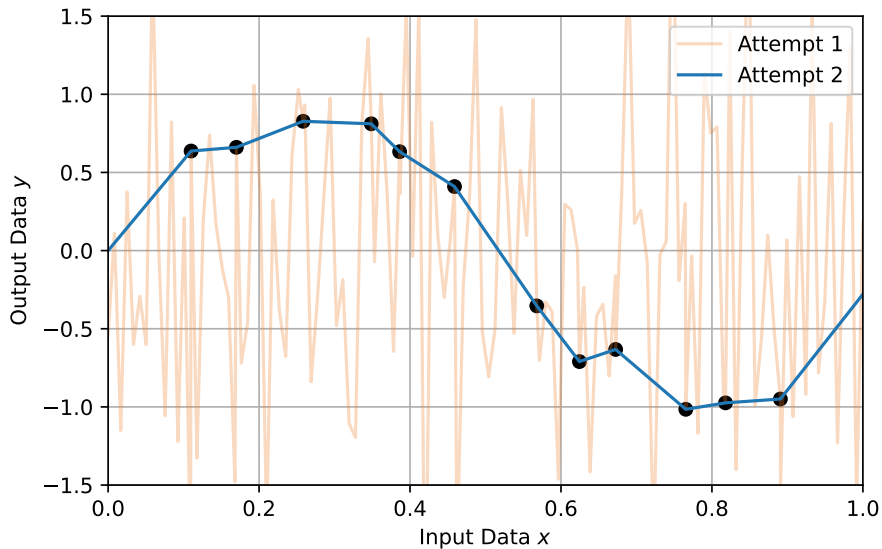
What happens between the dots?



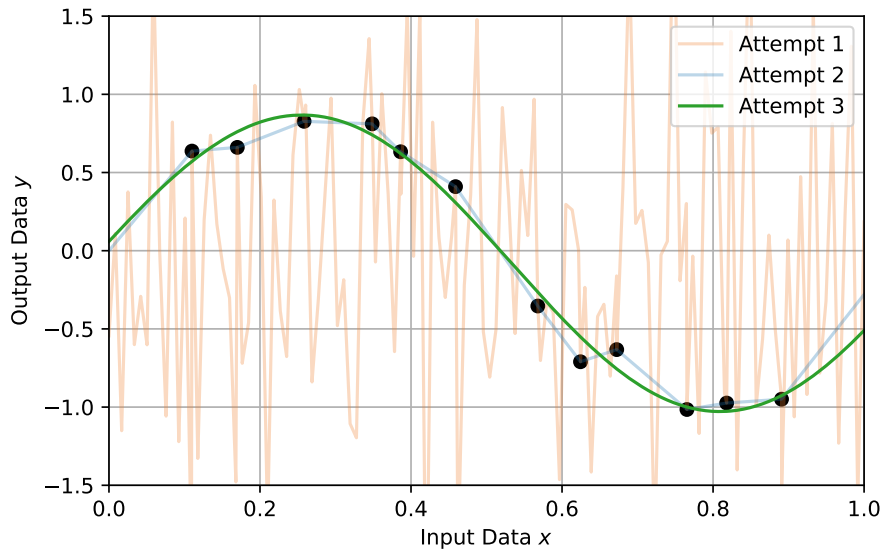
What happens between the dots?



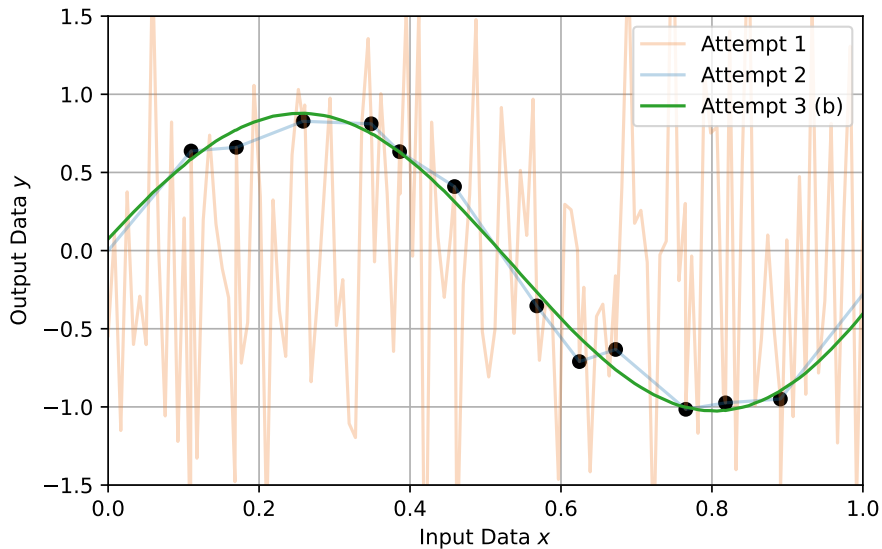
What happens between the dots?



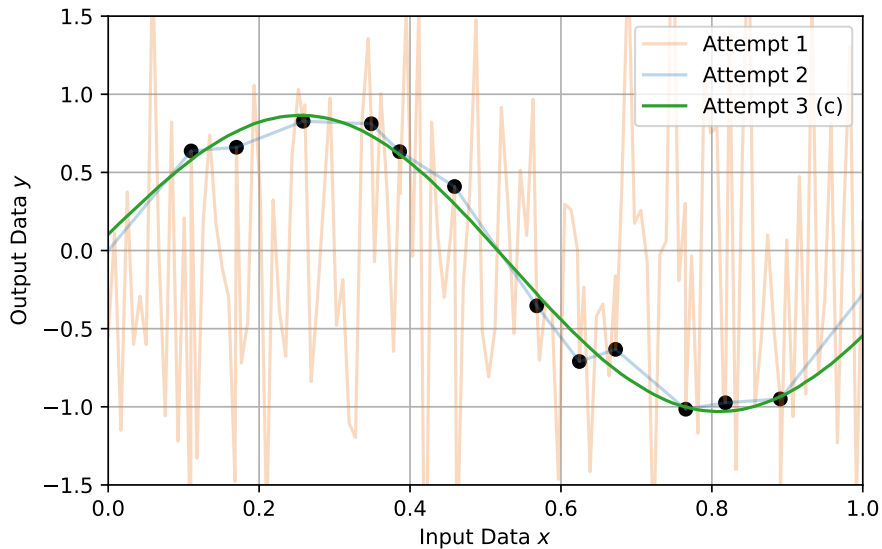
What happens between the dots?



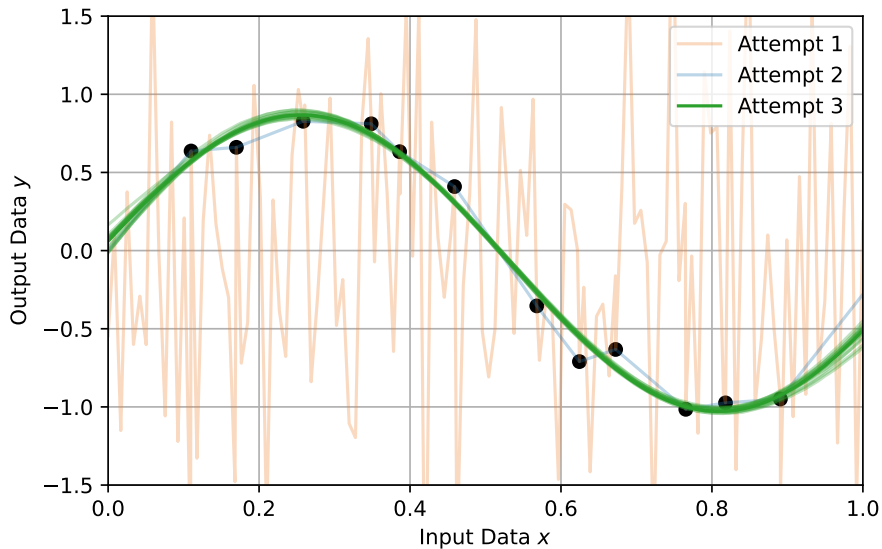
What happens between the dots?



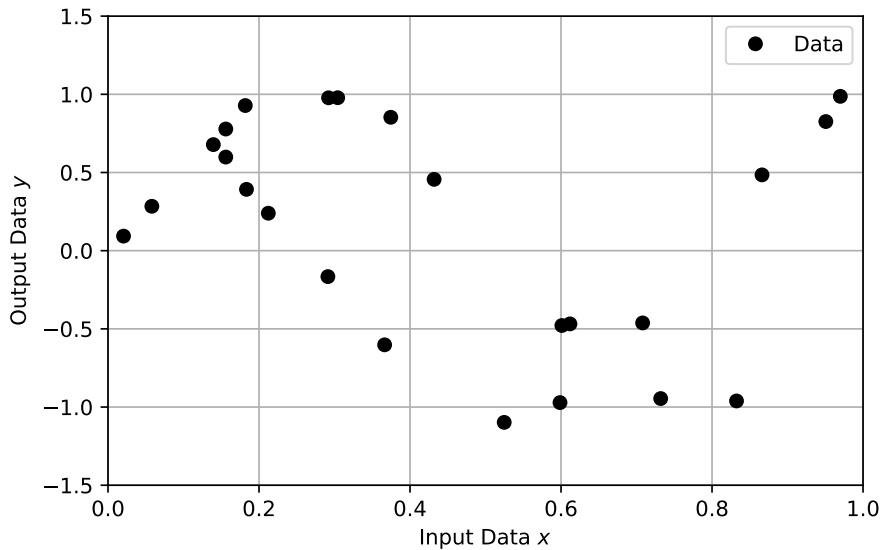
What happens between the dots?



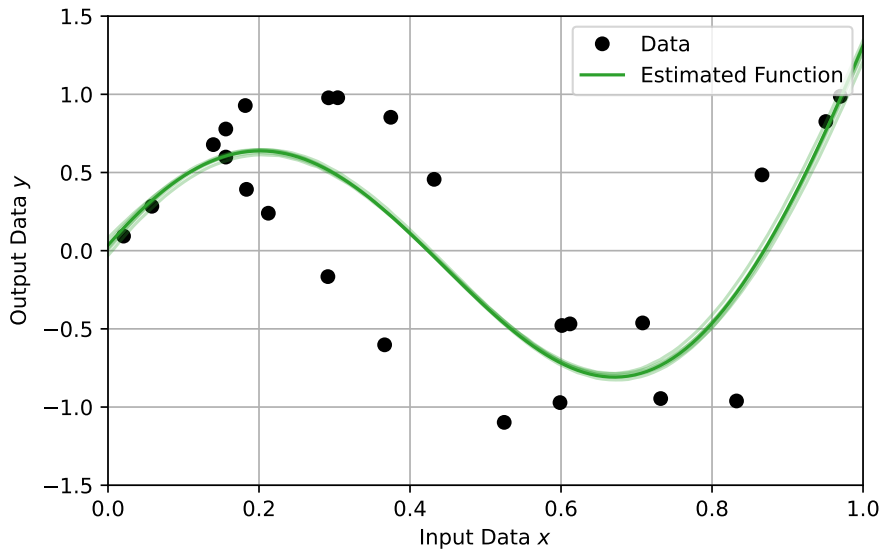
What happens between the dots?



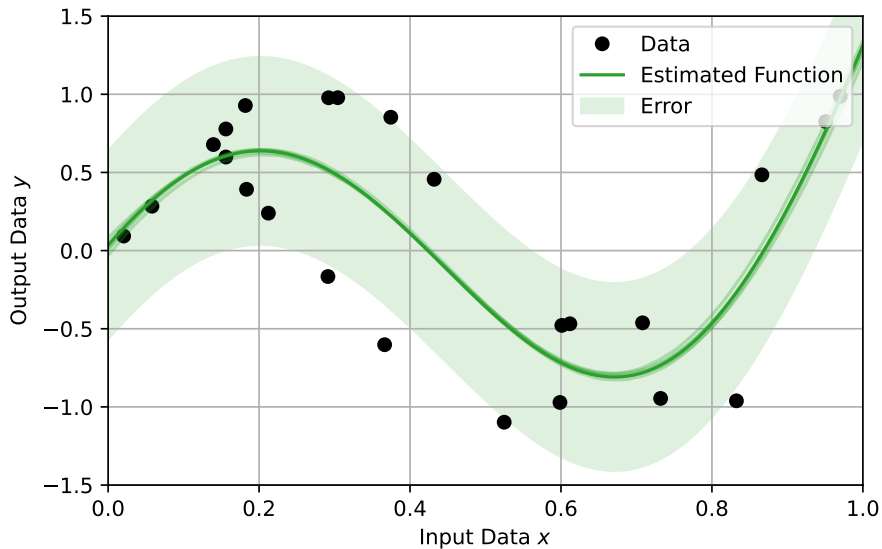
Ambiguity..



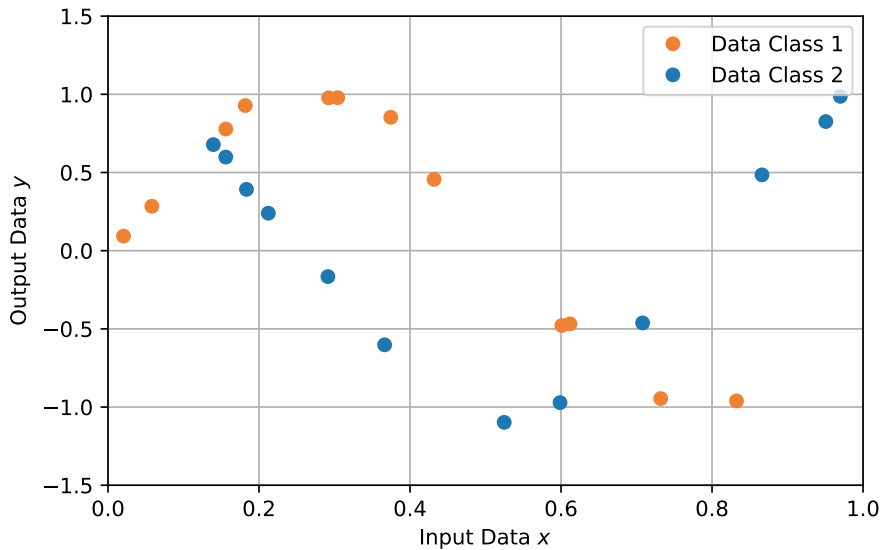
Ambiguity..



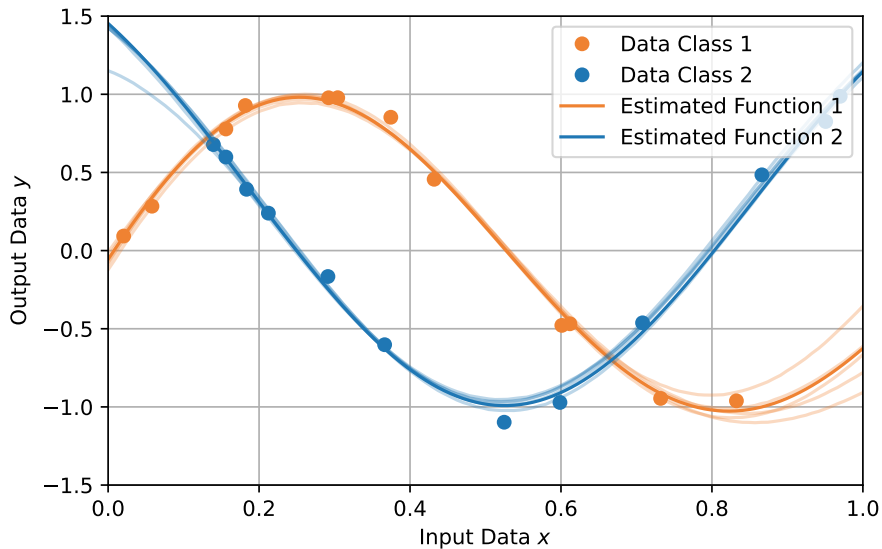
Ambiguity..



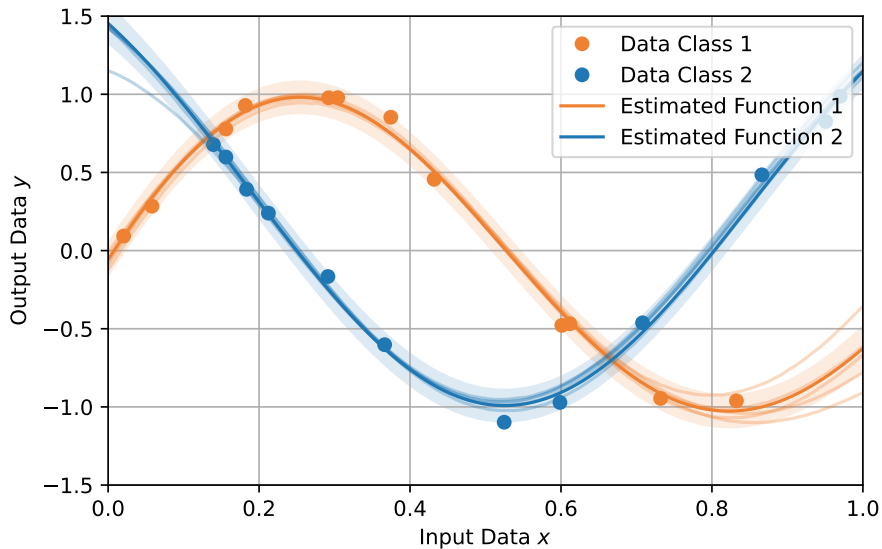
Ambiguity..



Ambiguity..

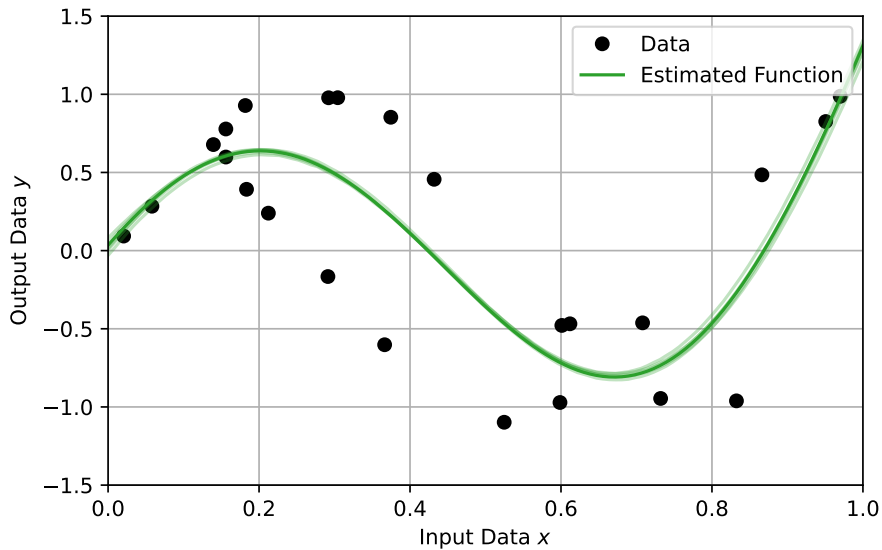


Ambiguity..

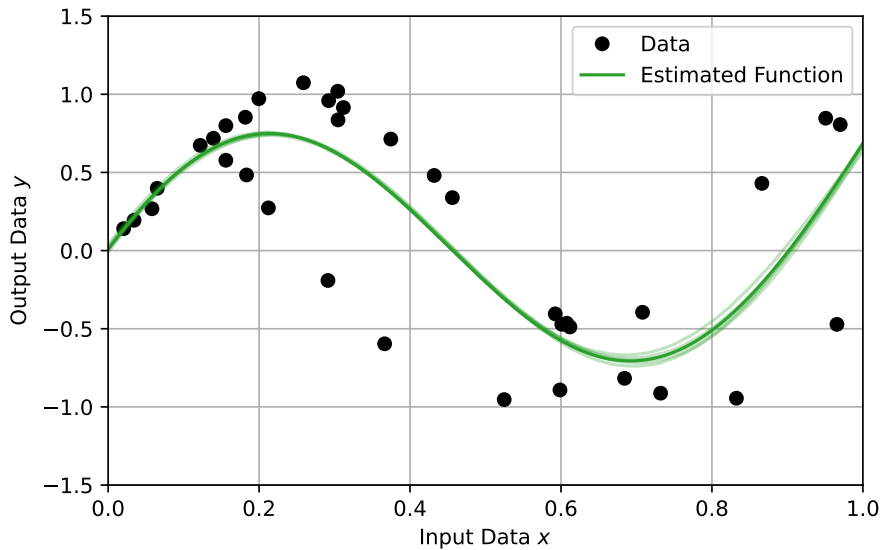


Average vs Worst Case..

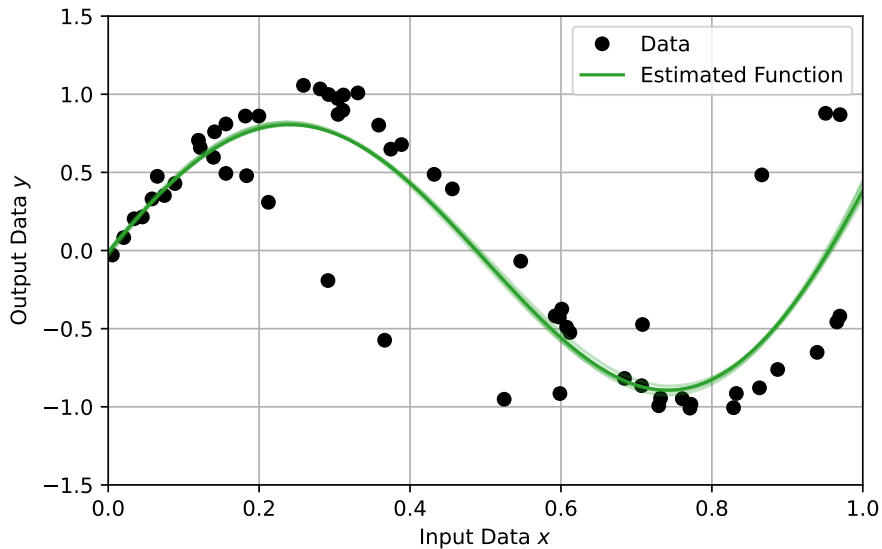
Average vs Worst Case..



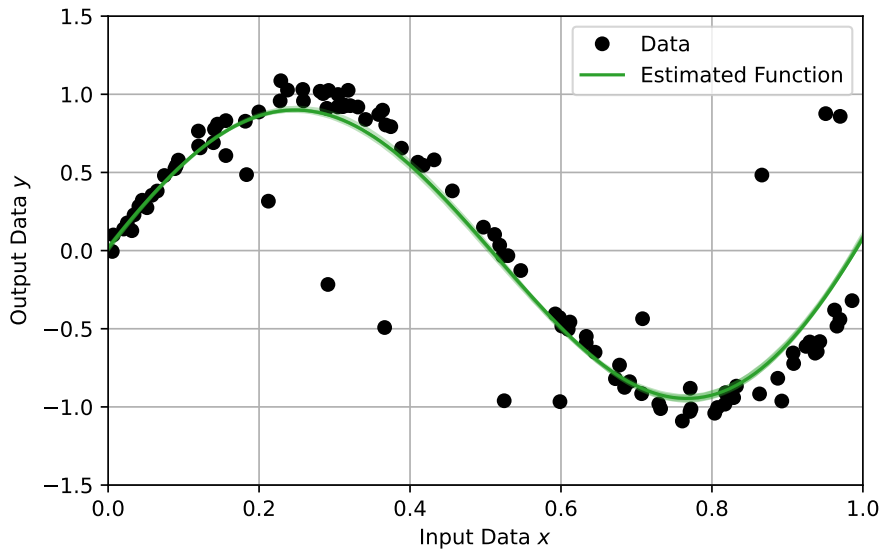
Average vs Worst Case..



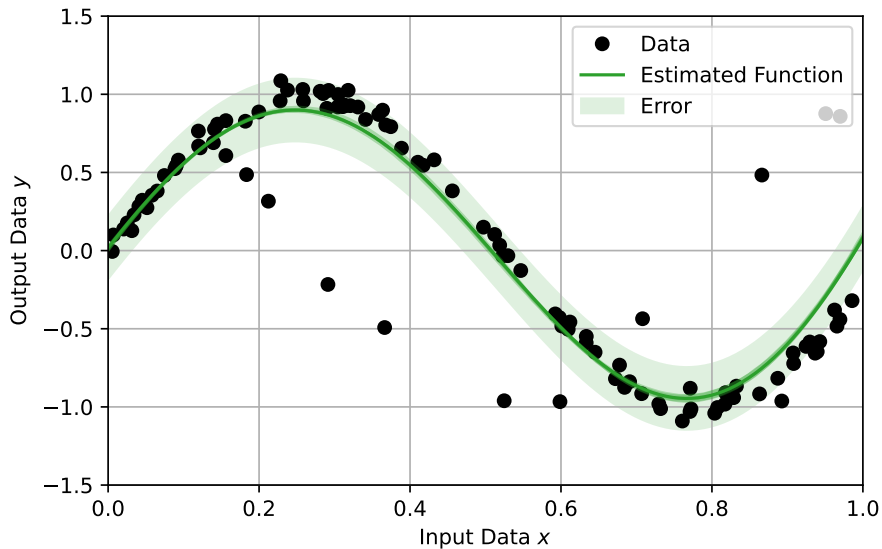
Average vs Worst Case..



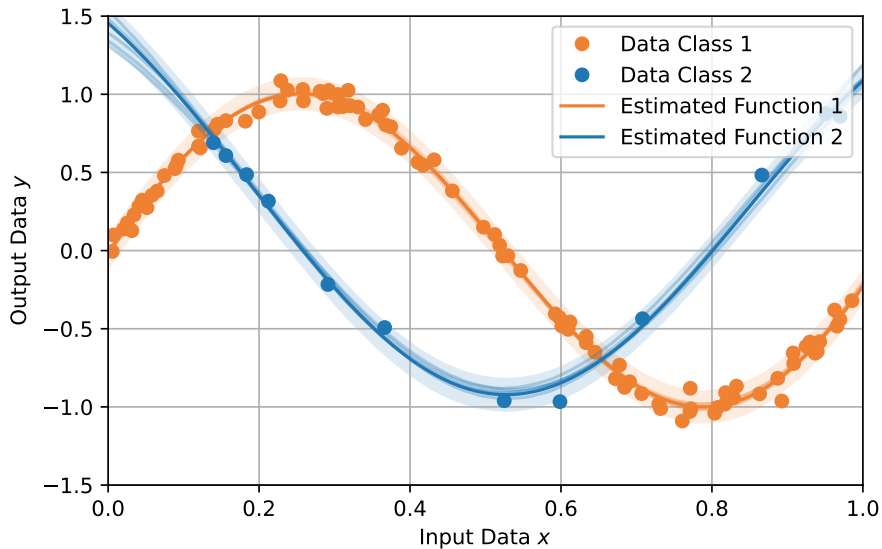
Average vs Worst Case..



Average vs Worst Case: Failure to model..

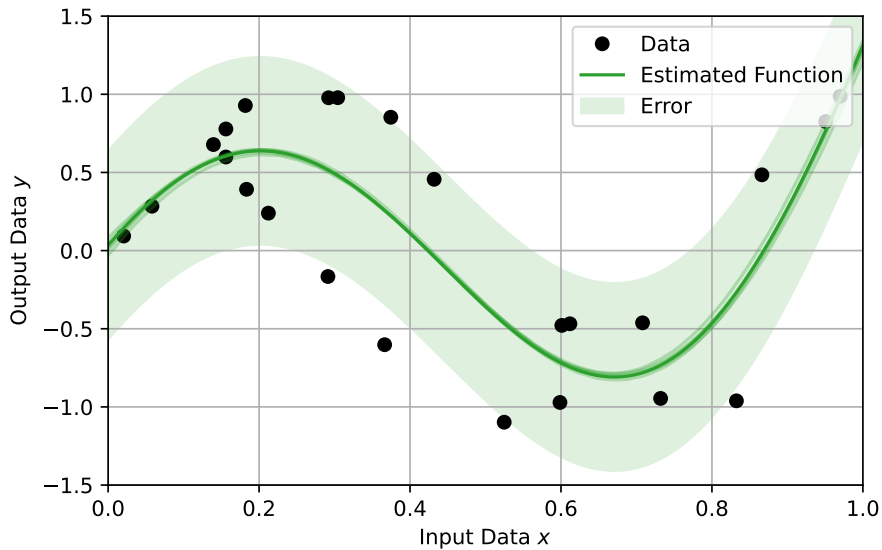


Average vs Worst Case: Explicitly accounting for imbalance..

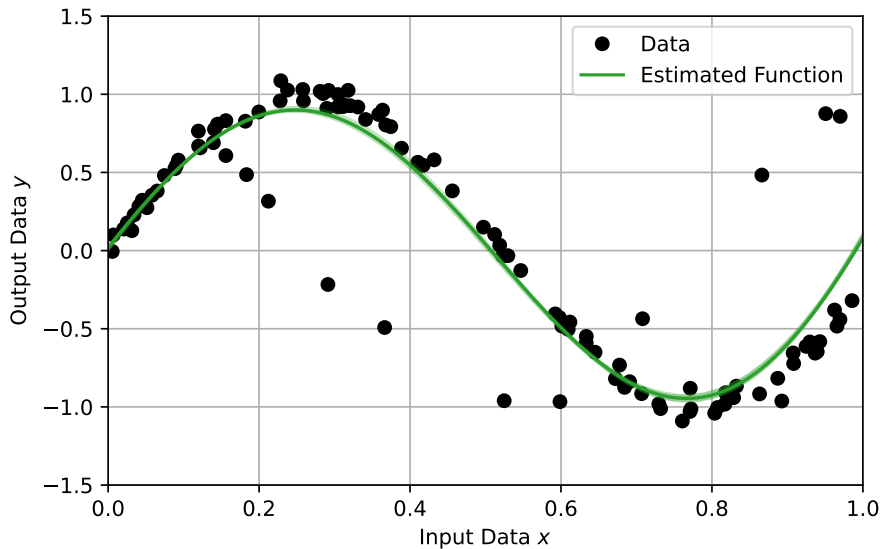


Story of Machine Learning (in three slides!)

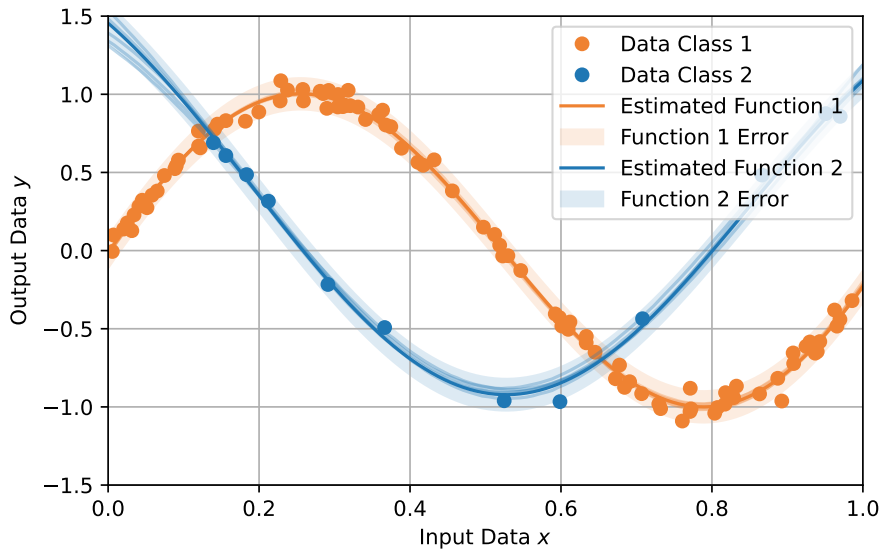
Story of ML: Past...



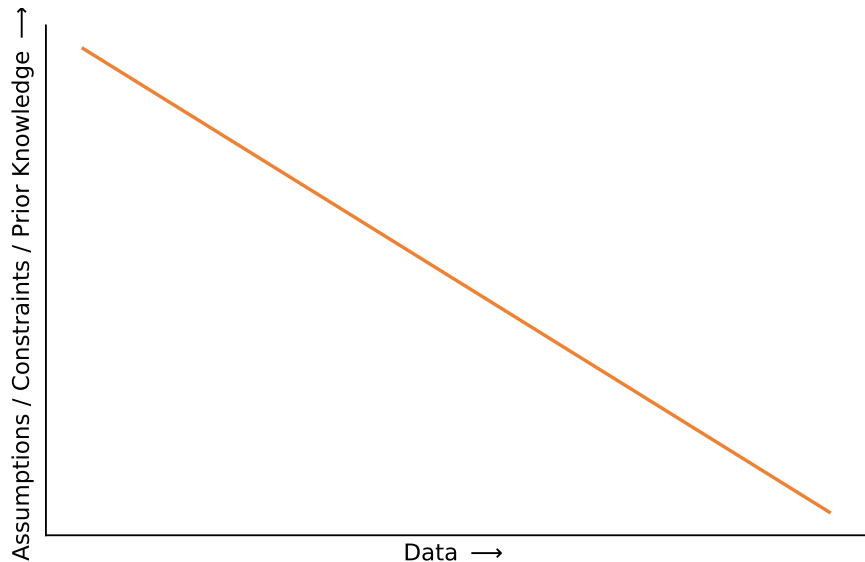
Story of ML: Present...



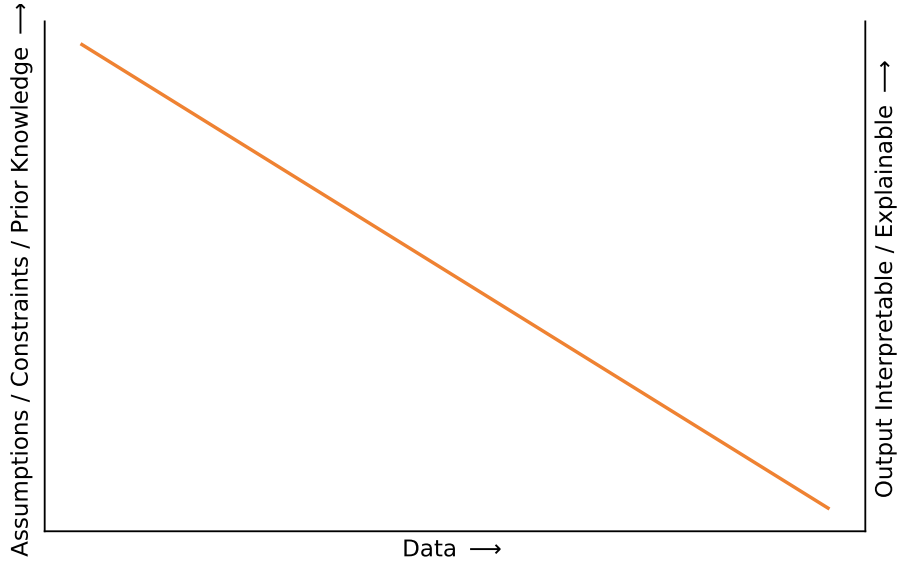
Story of ML: Future...



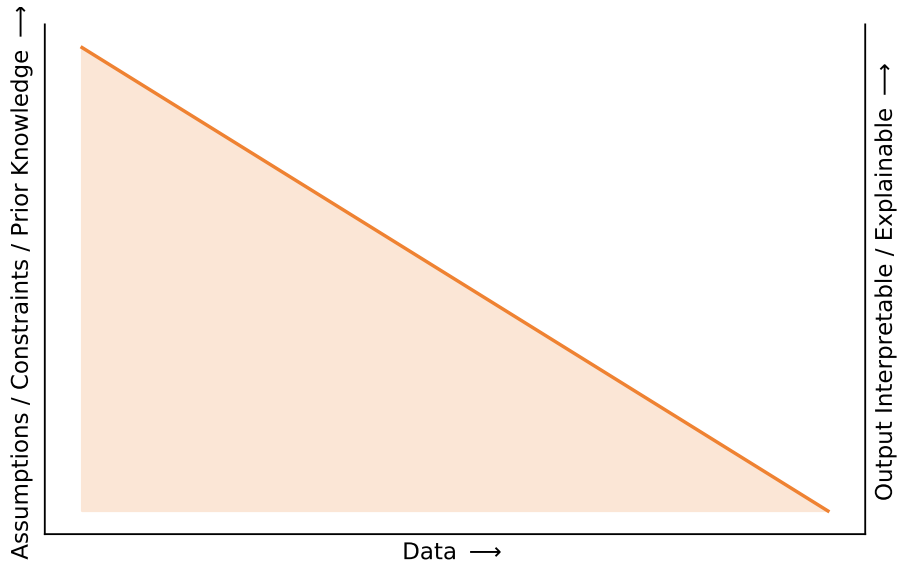
No free lunch



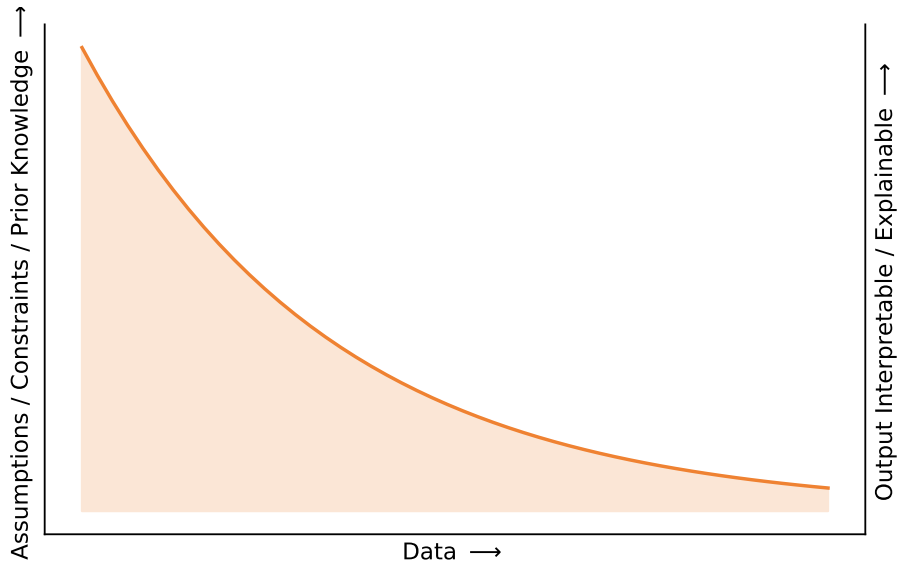
No free lunch



No free lunch



No free lunch (more realistic)



Why are we going to look at Bayesian methods?

“The Theory of probability is simply common sense reduced to calculus”

Pierre-Simon Laplace, 1749-1827

- We use Machine Learning to deal with the **unknown**
- Bayesian probability is the application of logic in the face of uncertainty

Why are we going to look at Bayesian methods?

“The Theory of probability is simply common sense reduced to calculus”

Pierre-Simon Laplace, 1749-1827

- We use Machine Learning to deal with the **unknown**
- Bayesian probability is the application of logic in the face of uncertainty
- Applications usually care about **Inverse Probability**..

Whirlwind Introduction to Inverse Probabilities

Overview

Motivation

No Free Lunch

Whirlwind Introduction to Inverse Probabilities

Model Selection

Evaluation

Bayesian Machine Learning: Simple Example

Why don't we do Model Selection in Machine Learning?

Conclusions

Monty Hall..



Monty Hall..



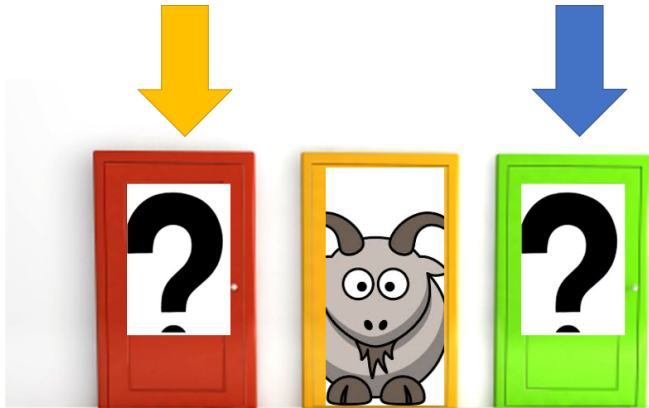
Monty Hall..



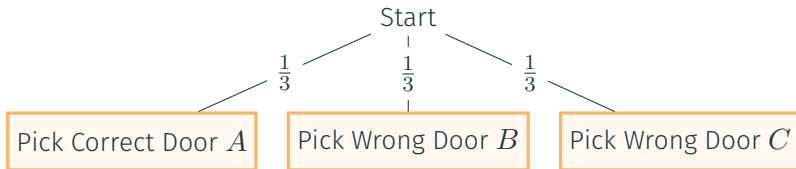
Monty Hall..



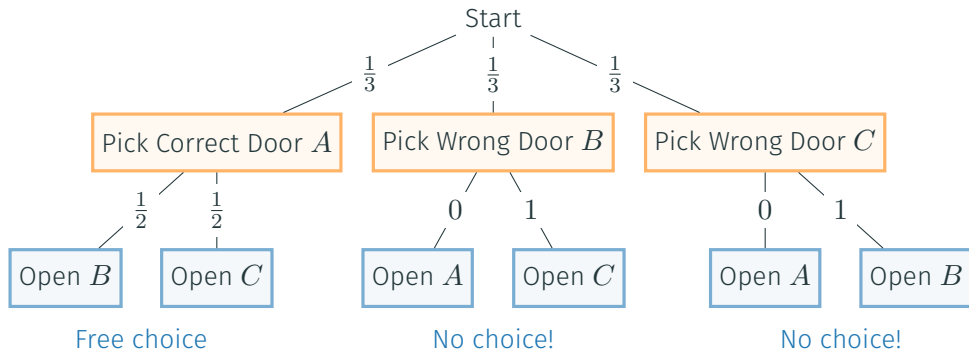
Monty Hall..



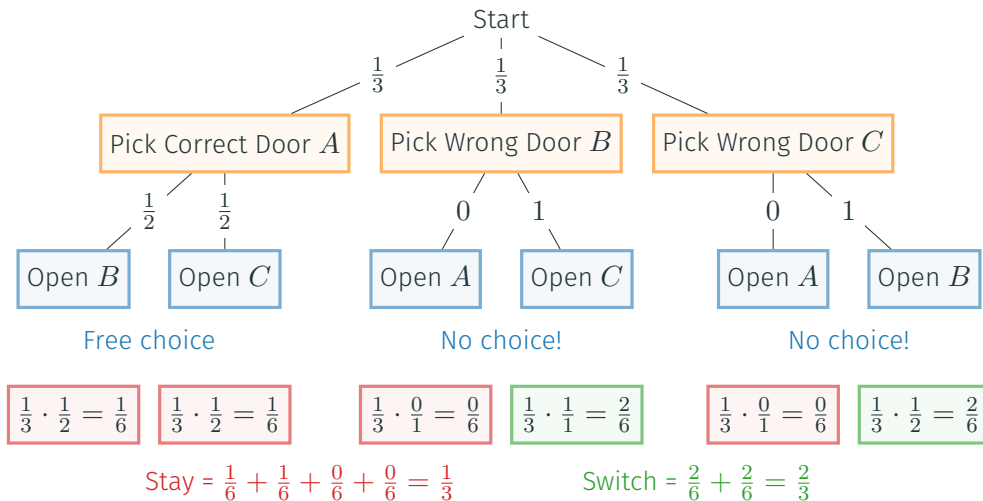
Monty Hall: Think of a tree diagram...



Monty Hall: Think of a tree diagram...



Monty Hall: Think of a tree diagram...



Monty Hall: How would we generate data (or simulate)?

Monty Hall: How would we generate data (or simulate)?

```
1 door_with_car = pick_random({1, 2, 3})
2 door_with_goat = {1, 2, 3} - door_with_car
3
4 door_picked = pick_random({1, 2, 3})
5
6 if door_picked == door_with_car:
7     door_to_open = pick_random(door_with_goat)
8 else:
9     door_to_open = door_with_goat - door_picked
```

Monty Hall: How would we generate data (or simulate)?

```
1 door_with_car = pick_random({1, 2, 3})           # 1/3 equal chance
2 door_with_goat = {1, 2, 3} - door_with_car
3
4 door_picked = pick_random({1, 2, 3})             # 1/3 equal chance
5
6 if door_picked == door_with_car:
7     door_to_open = pick_random(door_with_goat)   # 1 times in 3
8 else:
9     door_to_open = door_with_goat - door_picked  # 2 times in 3
```

Consider **Modelling** as a **Generative
Process**

The Rules of Probability

- Notation $p(a) = p(a = \mathcal{A})$ where a is a particular outcome chosen from the set of all possible outcomes $\mathcal{A}$

$p(A | B)$ means “probability of A being the case given that B occurs”

- Probabilities in the range $0 \rightarrow 1$
- $0 =$ impossible
- $1 =$ certain
- Sum over all possible outcomes must be 1

The Rules of Probability

The Sum Rule (Marginalisation)

$$p(A = a) = \sum_{b \in \mathbb{B}} p(A = a, B = b)$$

- If continuous, rather than discrete, use densities and

$$p(A = a) = \int_{\mathbb{R}} p(A = a, B = b) db$$

The Product Rule

$$p(A = a, B = b) = p(A = a \mid B = b) p(B = b) = p(B = b \mid A = a) p(A = a)$$

- Bayes' rule follows from these rules..
- Only consistent approach for probability as “degree of plausibility” (Cox)

Bayes' Rule

- From the product rule

$$p(A = a \mid B = b) = \frac{p(B = b \mid A = a) p(A = a)}{p(B = b)}$$

- We can also **condition** on other information $\mathcal{H}$

$$p(a \mid b, \mathcal{H}) = \frac{p(b \mid a, \mathcal{H}) p(a \mid \mathcal{H})}{p(b \mid \mathcal{H})}$$

- We give the parts of equation specific terms

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

Bayes' Rule

$$\text{Posterior Probability (after)} = \frac{\text{Likelihood (of event)} \times \text{Prior Probability (before)}}{\text{Evidence}}$$

Example of Bayes' Rule..

- Consider a legal trial..

$$\underbrace{p(\text{guilt} \mid \text{observations})}_{\text{Posterior}} = \frac{\overbrace{p(\text{observations} \mid \text{guilt})}^{\text{Likelihood}} \times \overbrace{p(\text{guilt})}^{\text{Prior}}}{\underbrace{p(\text{observations})}_{\text{Evidence}}}$$

Example of Bayes' Rule..

- Consider a legal trial..

$$\underbrace{p(\text{guilt} \mid \text{observations})}_{\text{Posterior}} = \frac{\overbrace{p(\text{observations} \mid \text{guilt})}^{\text{Likelihood}} \times \overbrace{p(\text{guilt})}^{\text{Prior}}}{\underbrace{\sum_{\text{guilt}} p(\text{observations} \mid \text{guilt}) p(\text{guilt})}_{\text{Evidence}}}$$

- The evidence is the sum of the top row over both the guilty ($\text{guilt} = 1$) and innocent ($\text{guilt} = 0$) cases.
- We might want to be careful about how we treat $p(\text{guilt})$.

Example of Bayes' Rule..

- Consider a legal trial..

$$\underbrace{p(\text{guilt} \mid \text{observations})}_{\text{Posterior}} = \frac{\overbrace{p(\text{observations} \mid \text{guilt})}^{\text{Likelihood}} \times \overbrace{p(\text{guilt})}^{\text{Prior}}}{\underbrace{\sum_{\text{guilt}} p(\text{observations} \mid \text{guilt}) p(\text{guilt})}_{\text{Evidence}}}$$

- The evidence is the sum of the top row over both the guilty ($\text{guilt} = 1$) and innocent ($\text{guilt} = 0$) cases.
- We might want to be careful about how we treat $p(\text{guilt})$.

Example of Bayes' Rule..

- Consider a legal trial..

$$\underbrace{p(\text{guilt} \mid \text{observations})}_{\text{Posterior}} = \frac{\overbrace{p(\text{observations} \mid \text{guilt})}^{\text{Likelihood}} \times \overbrace{p(\text{guilt})}^{\text{Prior}}}{\underbrace{\sum_{\text{guilt}} p(\text{observations} \mid \text{guilt}) p(\text{guilt})}_{\text{Evidence}}}$$

- The evidence is the sum of the top row over both the guilty ($\text{guilt} = 1$) and innocent ($\text{guilt} = 0$) cases.
- We might want to be careful about how we treat $p(\text{guilt})$.
- Consider that the jury has to return a verdict “*beyond all reasonable doubt*”..

Bayes' Rule with models and functions..

$$\underbrace{p(\text{functions} \mid \text{observed data})}_{\text{Posterior}} = \frac{\overbrace{p(\text{observed data} \mid \text{functions})}^{\text{Likelihood}} \times \overbrace{p(\text{functions})}^{\text{Prior}}}{\underbrace{p(\text{observed data})}_{\text{Evidence}}}$$

Bayes' Rule with models and functions..

$$\underbrace{p(\text{functions} \mid \text{observed data})}_{\text{Posterior}} = \frac{\overbrace{p(\text{observed data} \mid \text{functions})}^{\text{Likelihood}} \times \overbrace{p(\text{functions})}^{\text{Prior}}}{\underbrace{p(\text{observed data})}_{\text{Evidence}}}$$

$$\underbrace{p(f \mid \mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{p(\mathcal{D} \mid f)}^{\text{Likelihood}} \times \overbrace{p(f)}^{\text{Prior}}}{\underbrace{p(\mathcal{D})}_{\text{Evidence}}} \qquad p(\mathcal{D}) = \sum_f p(\mathcal{D} \mid f) p(f)$$

Data $\mathcal{D} = \{X, Y\}$, pairs of inputs $\{x_n\}$ and outputs $\{y_n\}$, and functions f

Bayes' Rule with models and functions..

$$\underbrace{p(\text{functions} \mid \text{observed data})}_{\text{Posterior}} = \frac{\overbrace{p(\text{observed data} \mid \text{functions})}^{\text{Likelihood}} \times \overbrace{p(\text{functions})}^{\text{Prior}}}{\underbrace{p(\text{observed data})}_{\text{Evidence}}}$$

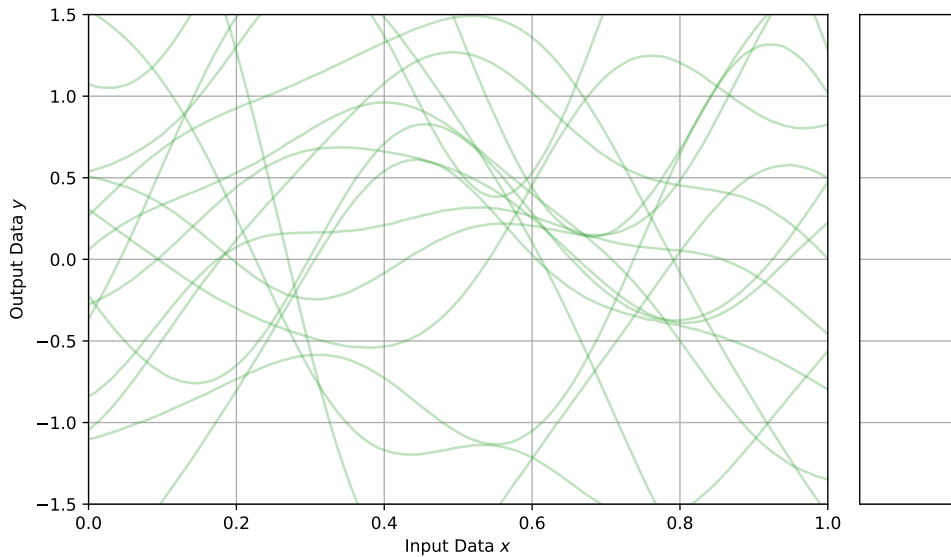
$$\underbrace{p(f \mid \mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{p(\mathcal{D} \mid f)}^{\text{Likelihood}} \times \overbrace{p(f)}^{\text{Prior}}}{\underbrace{p(\mathcal{D})}_{\text{Evidence}}} \quad p(\mathcal{D}) = \sum_f p(\mathcal{D} \mid f) p(f)$$

Data $\mathcal{D} = \{X, Y\}$, pairs of inputs $\{x_n\}$ and outputs $\{y_n\}$, and functions f

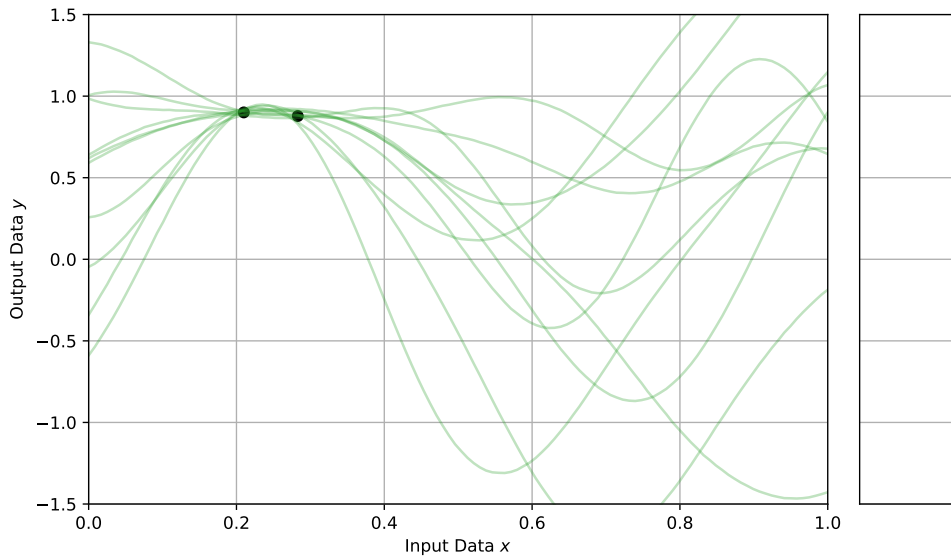
Average over functions to predict unknown output y^* for a new input x^* :

$$p(y^* \mid x^*, \mathcal{D}) = \sum_f p(y^* \mid x^*, f) p(f \mid \mathcal{D})$$

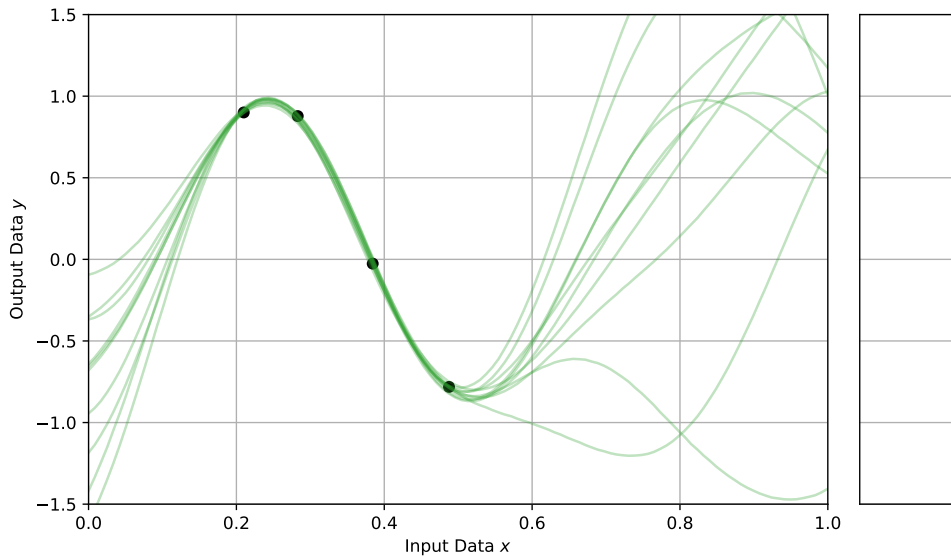
Prior over functions...



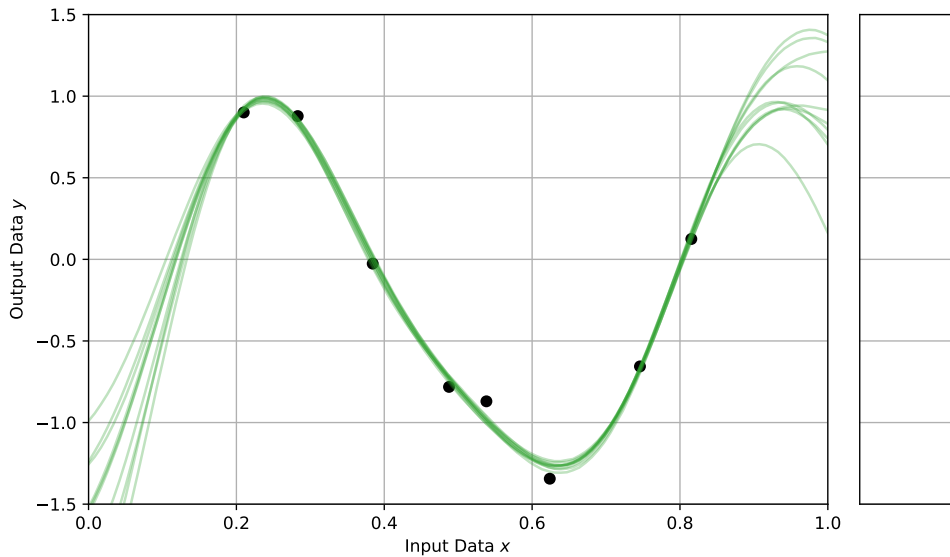
Combine prior with data...



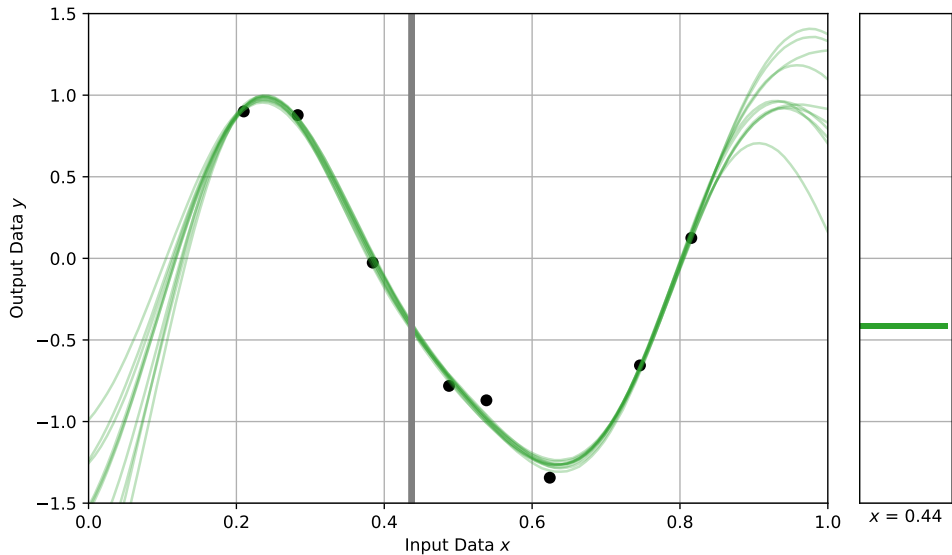
Combine prior with data...



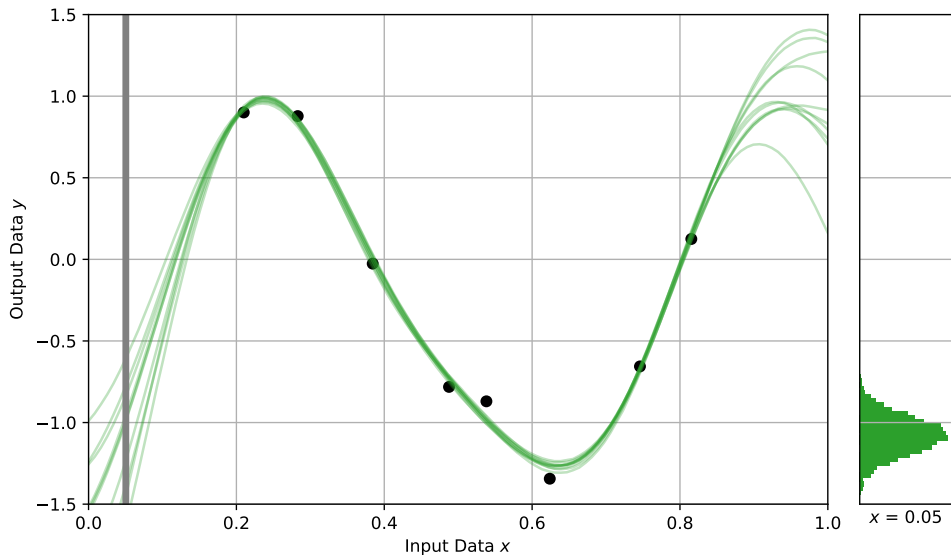
Combine prior with data...



Average over functions to predict...



Averaging over functions gives us (Epistemic) Uncertainty!



Bayes' Rule with models and parameters..

$$\underbrace{p(\text{params} \mid \text{obs. data})}_{\text{Posterior}} = \frac{\overbrace{p(\text{obs. data} \mid \text{params})}^{\text{Likelihood}} \times \overbrace{p(\text{params})}^{\text{Prior}}}{\underbrace{p(\text{obs. data})}_{\text{Evidence}}}$$

Bayes' Rule with models and parameters..

$$\underbrace{p(\text{params} \mid \text{obs. data})}_{\text{Posterior}} = \frac{\overbrace{p(\text{obs. data} \mid \text{params})}^{\text{Likelihood}} \times \overbrace{p(\text{params})}^{\text{Prior}}}{\underbrace{p(\text{obs. data})}_{\text{Evidence}}}$$

$$\underbrace{p(w \mid \mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{p(\mathcal{D} \mid w)}^{\text{Likelihood}} \times \overbrace{p(w)}^{\text{Prior}}}{\underbrace{p(\mathcal{D})}_{\text{Evidence}}}$$

Data $\mathcal{D} = X, Y$, pairs of inputs x_n and outputs y_n

Bayes' Rule with models and parameters..

$$\underbrace{p(\text{params} \mid \text{obs. data})}_{\text{Posterior}} = \frac{\overbrace{p(\text{obs. data} \mid \text{params})}^{\text{Likelihood}} \times \overbrace{p(\text{params})}^{\text{Prior}}}{\underbrace{p(\text{obs. data})}_{\text{Evidence}}}$$

$$\underbrace{p(w \mid \mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{p(\mathcal{D} \mid w)}^{\text{Likelihood}} \times \overbrace{p(w)}^{\text{Prior}}}{\underbrace{p(\mathcal{D})}_{\text{Evidence}}} \quad p(\mathcal{D}) = \sum_w p(\mathcal{D} \mid w) p(w)$$

Data $\mathcal{D} = X, Y$, pairs of inputs x_n and outputs y_n

Prediction of output y^* for a new input x^* :

$$p(y^* \mid x^*, \mathcal{D}) = \sum_w p(y^* \mid x^*, w) p(w \mid \mathcal{D})$$

New thought process; no longer “Find the best parameters”, now “Find all the parameters that agree with the data”..

Model Selection

Overview...

Overview

Motivation

No Free Lunch

Whirlwind Introduction to Inverse Probabilities

Model Selection

Evaluation

Bayesian Machine Learning: Simple Example

Why don't we do Model Selection in Machine Learning?

Conclusions

Model selection (“Which model should I use?”)

- How much data do we need?

Model selection (“Which model should I use?”)

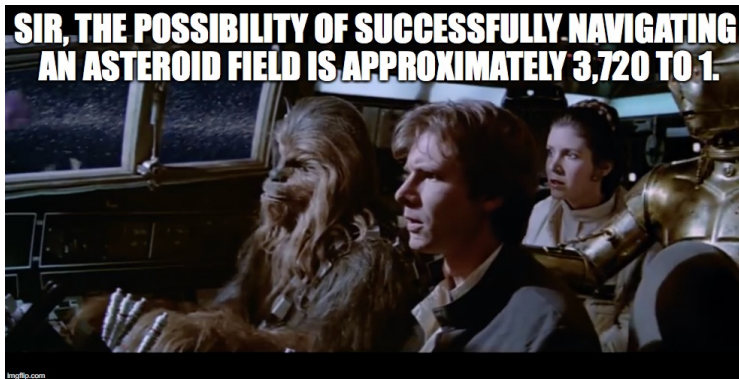
- How much data do we need?
- Might not be the right question..

Model selection (“Which model should I use?”)

- How much data do we need?
- Might not be the right question..
- What can we actually say?

Model selection (“Which model should I use?”)

- How much data do we need?
- Might not be the right question..
- What can we actually say? **The odds!**



Model selection (“Which model should I use?”)

- How much data do we need?
- Might not be the right question..
- What can we actually say? **The odds!**



Science (and Machine Learning) cannot
prove things to be true via data

Science (and Machine Learning) cannot
prove things to be true via data
we can only demonstrate that things
are inconsistent with data

Model selection illustration: Gravity!



Stable Diffusion: *“Drop cannonball and orange off the leaning tower of Pisa.”*

Model selection illustration: Gravity!

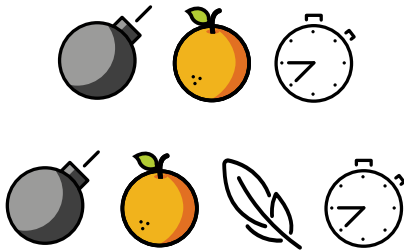


Stable Diffusion: *“Drop cannonball and orange off the leaning tower of Pisa.”*

Model selection illustration: Gravity!



Stable Diffusion: *“Drop cannonball and orange off the leaning tower of Pisa.”*



Model selection illustration: Gravity!



Stable Diffusion: *“Drop cannonball and orange off the leaning tower of Pisa.”*



Model selection illustration: Gravity!



Stable Diffusion: *“Drop cannonball and orange off the leaning tower of Pisa.”*



Apollo 15 Hammer-Feather Drop



NASASolarSystem
14.9K subscribers

Subscribe

4.8K



Share



576K views · 8 years ago

At the end of the last Apollo 15 moon walk, Commander David Scott (pictured above) performed a live demonstration for the television cameras. He held out a geologic hammer and a feather and dropped them at the same time. Because they were essentially in a vacuum, there w ...more

Bayes' Rule for model selection..

$$\underbrace{p(w \mid \mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{p(\mathcal{D} \mid w)}^{\text{Likelihood}} \times \overbrace{p(w)}^{\text{Prior}}}{\underbrace{p(\mathcal{D})}_{\text{Evidence}}}$$

Data $\mathcal{D} = \{X, Y\}$, input/output pairs, and parameters w

Bayes' Rule for model selection..

$$\underbrace{p(w \mid \mathcal{D}, \mathcal{M} = m)}_{\text{Posterior under model}} = \frac{\overbrace{p(\mathcal{D} \mid w, \mathcal{M} = m)}^{\text{Likelihood under model}} \times \overbrace{p(w, \mathcal{M} = m)}^{\text{Prior}}}{\underbrace{p(\mathcal{D} \mid \mathcal{M} = m)}_{\text{Evidence for model}}}$$

Data $\mathcal{D} = \{X, Y\}$, input/output pairs, and parameters w for Model $\mathcal{M} = m$

Bayes' Rule for model selection..

$$\underbrace{p(w \mid \mathcal{D}, \mathcal{M} = m)}_{\text{Posterior under model}} = \frac{\overbrace{p(\mathcal{D} \mid w, \mathcal{M} = m)}^{\text{Likelihood under model}} \times \overbrace{p(w, \mathcal{M} = m)}^{\text{Prior}}}{\underbrace{p(\mathcal{D} \mid \mathcal{M} = m)}_{\text{Evidence for model}}}$$

Data $\mathcal{D} = \{X, Y\}$, input/output pairs, and parameters w for Model $\mathcal{M} = m$

$$\underbrace{p(\mathcal{M} = m \mid \mathcal{D})}_{\text{Posterior for model}} = \frac{\overbrace{p(\mathcal{D} \mid \mathcal{M} = m)}^{\text{Evidence for model}} \times \overbrace{p(\mathcal{M} = m)}^{\text{Prior for model}}}{\underbrace{p(\mathcal{D})}_{\text{Data}}}$$

Bayes' Rule for model selection..

$$\underbrace{p(w \mid \mathcal{D}, \mathcal{M} = m)}_{\text{Posterior under model}} = \frac{\overbrace{p(\mathcal{D} \mid w, \mathcal{M} = m)}^{\text{Likelihood under model}} \times \overbrace{p(w, \mathcal{M} = m)}^{\text{Prior}}}{\underbrace{p(\mathcal{D} \mid \mathcal{M} = m)}_{\text{Evidence for model}}}$$

Data $\mathcal{D} = \{X, Y\}$, input/output pairs, and parameters w for Model $\mathcal{M} = m$

$$\underbrace{p(\mathcal{M} = m \mid \mathcal{D})}_{\text{Posterior for model}} = \frac{\overbrace{p(\mathcal{D} \mid \mathcal{M} = m)}^{\text{Evidence for model}} \times \overbrace{p(\mathcal{M} = m)}^{\text{Prior for model}}}{\underbrace{p(\mathcal{D})}_{\text{Data}}}$$

If prior over models is equal, we compare via the **Evidence for the Model**: $p(\mathcal{D} \mid \mathcal{M} = m)$

Model selection example

Fitting polynomial models to data under Gaussian noise, $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$:

Model 1 : $y_n = a_0 + a_1x_n + \varepsilon_n$

Model 2 : $y_n = a_0 + a_1x_n + a_2x^2 + \varepsilon_n$

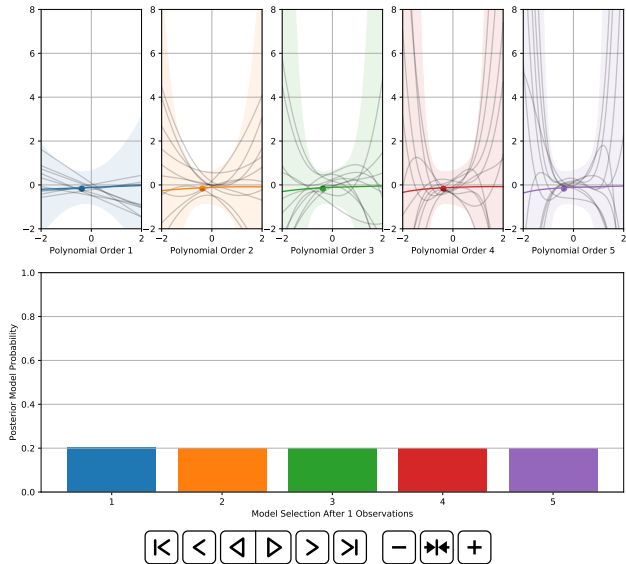
Model 3 : $y_n = a_0 + a_1x_n + a_2x^2 + a_3x^3 + \varepsilon_n$

Model 4 : $y_n = a_0 + a_1x_n + a_2x^2 + a_3x^3 + a_4x^4 + \varepsilon_n$

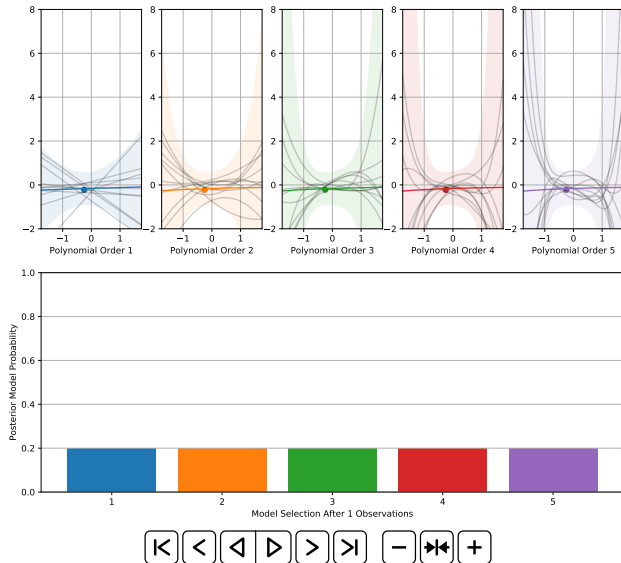
Model 5 : $y_n = a_0 + a_1x_n + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5 + \varepsilon_n$

Parameters $w_m = [a_0, \dots, a_m]$ for model m , where $m \in [1, \dots, 5]$.

Model selection example



Model selection example (more noise)



Evaluation

Overview...

Overview

Motivation

No Free Lunch

Whirlwind Introduction to Inverse Probabilities

Model Selection

Evaluation

Bayesian Machine Learning: Simple Example

Why don't we do Model Selection in Machine Learning?

Conclusions

Evaluation in Machine Learning

We have made a cool new model with great advances..

- How do we evaluate empirical performance?
- What do we care about?

Evaluation in Machine Learning

We have made a cool new model with great advances..

- How do we evaluate empirical performance?
- What do we care about?

We probably want:

- Fair comparisons (comparing methods)
- Useful comparisons (what can we learn)
- Understand limitations (how confident should we be)

Evaluation in Machine Learning

We have made a cool new model with great advances..

- How do we evaluate empirical performance?
- What do we care about?

We probably want:

- Fair comparisons (comparing methods)
- Useful comparisons (what can we learn)
- Understand limitations (how confident should we be)
- **How can we make it better if we don't know where it fails?!**

Illustrative example: Which method is best?

Model name	RMSE Mean ↓
MD2 Boot+Log	3.850
MD2 Boot+Self	3.795
Diagonal	4.000
SUPN Boot+Log	4.071
SUPN Boot+Self	4.091

Illustrative example: Which method is best?

Model name	RMSE Mean ↓
MD2 Boot+Log	3.850
MD2 Boot+Self	3.795
Diagonal	4.000
SUPN Boot+Log	4.071
SUPN Boot+Self	4.091

- We are showing the Root of the Mean Squared Error
- We want lower scores so the MD2 model is the best?

Illustrative example: Which method is best?

Model name	RMSE Mean ↓
MD2 Boot+Log	3.850 (1.370)
MD2 Boot+Self	3.795 (1.397)
Diagonal	4.000 (1.457)
SUPN Boot+Log	4.071 (1.489)
SUPN Boot+Self	4.091 (1.442)

Illustrative example: Which method is best?

Model name	RMSE Mean ↓
MD2 Boot+Log	3.850 (1.370)
MD2 Boot+Self	3.795 (1.397)
Diagonal	4.000 (1.457)
SUPN Boot+Log	4.071 (1.489)
SUPN Boot+Self	4.091 (1.442)

- We now have standard errors (i.e. the standard deviation)
- Does this change things?

Illustrative example: Which method is best?

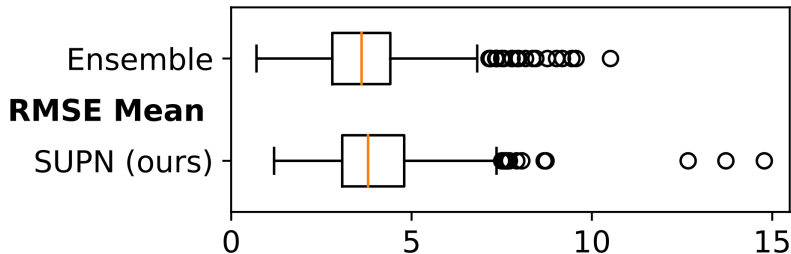


Figure 2. Box plot illustrating the strong distribution overlap between the original ensemble and the trained SUPN model for Boot+Log RMSE mean.

Illustrative example: Which method is best?

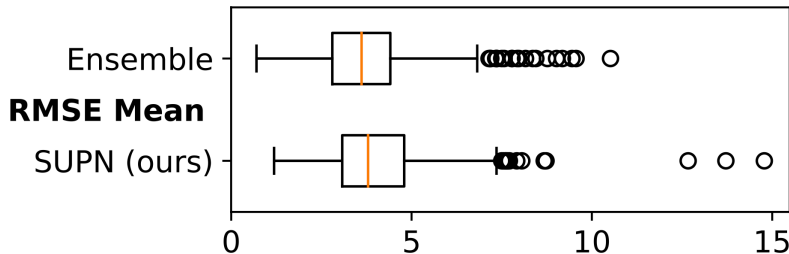


Figure 2. Box plot illustrating the strong distribution overlap between the original ensemble and the trained SUPN model for Boot+Log RMSE mean.

- Looking at the **distributions** reveals very similar performance
- The SUPN mean is skewed by only 3 outliers!

Important slide acknowledgements!

Important slide acknowledgements!

Illustrations taken from the excellent new text book from Simon Prince:

Understanding Deep Learning, Simon J.D. Prince, MIT Press

Final draft available on the website:

<https://udlbook.github.io/udlbook/>

Important slide acknowledgements!

Illustrations taken from the excellent new text book from Simon Prince:

Understanding Deep Learning, Simon J.D. Prince, MIT Press

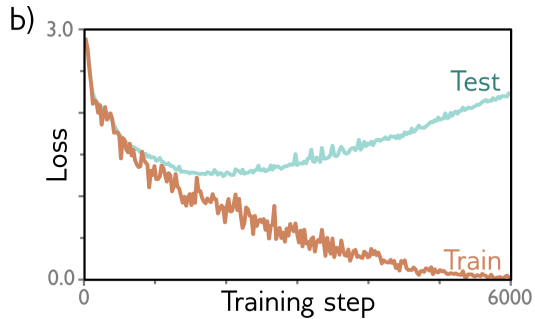
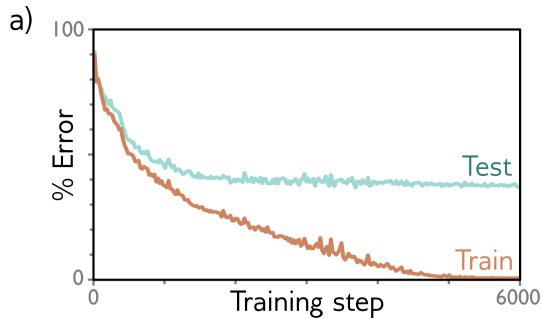
Final draft available on the website:

<https://udlbook.github.io/udlbook/>

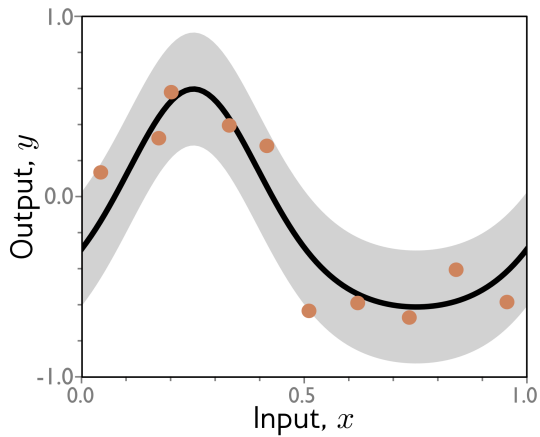
I strongly recommend you take a look!

Evaluation: Training vs Test

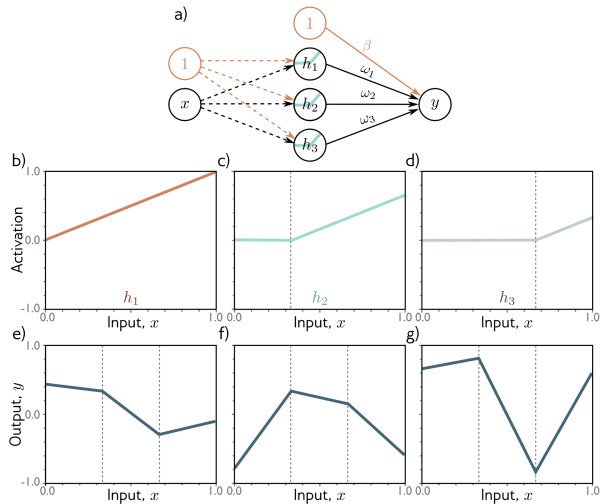
Evaluation: Training vs Test



Evaluation: Regression example



Evaluation: Toy regression model



Evaluation: Noise, Bias and Variance

$$y_n = f(x_n; \phi) + \varepsilon_n, \quad n = 1 \dots N, \quad \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$$

Evaluation: Noise, Bias and Variance

$$y_n = f(x_n; \phi) + \varepsilon_n, \quad n = 1 \dots N, \quad \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$$

$$y(x) = f(x; \phi) + \varepsilon$$

$$\mu(x) = \mathbb{E}_y[y(x)] = \int y(x) p(y \mid x) dy$$

Evaluation: Noise, Bias and Variance

$$y_n = f(x_n; \phi) + \varepsilon_n, \quad n = 1 \dots N, \quad \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$$

$$y(x) = f(x; \phi) + \varepsilon$$

$$\mu(x) = \mathbb{E}_y[y(x)] = \int y(x) p(y | x) dy$$

Now consider a least squares (L2) loss function:

$$\begin{aligned} \mathcal{L}(x, \phi) &= (f(x; \phi) - y(x))^2 \\ &= ((f(x; \phi) - \mu(x)) + (\mu(x) - y(x)))^2 \\ &= (f(x; \phi) - \mu(x))^2 + 2(f(x; \phi) - \mu(x)) (\mu(x) - y(x)) + (\mu(x) - y(x))^2 \end{aligned}$$

Evaluation: Noise, Bias and Variance

$$y_n = f(x_n; \phi) + \varepsilon_n, \quad n = 1 \dots N, \quad \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$$

$$y(x) = f(x; \phi) + \varepsilon$$

$$\mu(x) = \mathbb{E}_y[y(x)] = \int y(x) p(y | x) dy$$

Now consider a least squares (L2) loss function:

$$\begin{aligned} \mathcal{L}(x, \phi) &= (f(x; \phi) - y(x))^2 \\ &= ((f(x; \phi) - \mu(x)) + (\mu(x) - y(x)))^2 \\ &= (f(x; \phi) - \mu(x))^2 + 2(f(x; \phi) - \mu(x)) (\mu(x) - y(x)) + (\mu(x) - y(x))^2 \\ &\Rightarrow \mathbb{E}_y[\mathcal{L}(x, \phi)] = (f(x; \phi) - \mu(x))^2 + \sigma^2 \end{aligned}$$

Evaluation: Noise, Bias and Variance

$$\mathbb{E}_y[\mathcal{L}(x, \phi)] = (f(x; \phi) - \mu(x))^2 + \underbrace{\sigma^2}_{\text{noise}}$$

We have partitioned the expected loss into two terms, the second is some **irreducible noise** that comes with the observations (e.g. sensor noise).

Evaluation: Noise, Bias and Variance

$$\mathbb{E}_y[\mathcal{L}(x, \phi)] = (f(x; \phi) - \mu(x))^2 + \underbrace{\sigma^2}_{\text{noise}}$$

We have partitioned the expected loss into two terms, the second is some **irreducible noise** that comes with the observations (e.g. sensor noise).

So far we have ignored the fact that we actually estimate parameters from a **sampled dataset** $\mathcal{D} = \{x_n, y_n\}_{n=1}^N$

$$\begin{aligned} f_\mu(x) &:= \mathbb{E}_{\mathcal{D}}[f(x; \phi(\mathcal{D}))] \\ \Rightarrow (f(x; \phi) - \mu(x))^2 &= ((f(x; \phi_{\mathcal{D}}) - f_\mu(x)) + (f_\mu(x) - \mu(x)))^2 \\ \Rightarrow \mathbb{E}_{\mathcal{D}}[(f(x; \phi) - \mu(x))^2] &= \mathbb{E}_{\mathcal{D}}[(f(x; \phi_{\mathcal{D}}) - f_\mu(x))^2] + (f_\mu(x) - \mu(x))^2 \end{aligned}$$

Evaluation: Noise, Bias and Variance

Therefore, if we take expectations over datasets, our expected loss comprises three terms:

$$\mathbb{E}_{\mathcal{D}}[\mathbb{E}_y[\mathcal{L}(x, \phi)]] = \underbrace{\mathbb{E}_{\mathcal{D}}[(f(x; \phi_{\mathcal{D}}) - f_{\mu}(x))^2]}_{\text{variance}} + \underbrace{(f_{\mu}(x) - \mu(x))^2}_{\text{bias}} + \underbrace{\sigma^2}_{\text{noise}}$$

Note: more complex for models other than least squares...

Evaluation: Noise, Bias and Variance

Therefore, if we take expectations over datasets, our expected loss comprises three terms:

$$\mathbb{E}_{\mathcal{D}}[\mathbb{E}_y[\mathcal{L}(x, \phi)]] = \underbrace{\mathbb{E}_{\mathcal{D}}[(f(x; \phi_{\mathcal{D}}) - f_{\mu}(x))^2]}_{\text{variance}} + \underbrace{(f_{\mu}(x) - \mu(x))^2}_{\text{bias}} + \underbrace{\sigma^2}_{\text{noise}}$$

Note: more complex for models other than least squares...

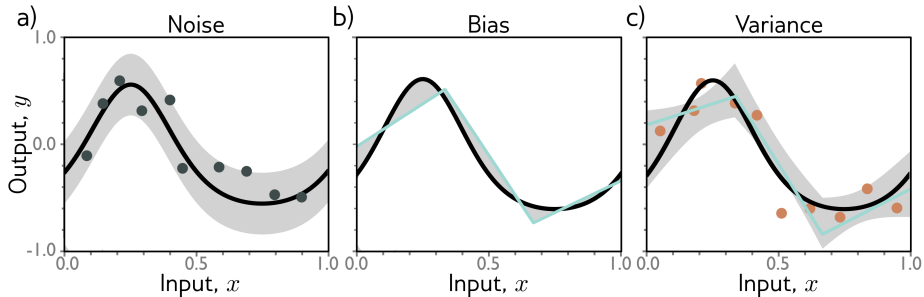
Noise Error in measurements (e.g. sensor noise, missing data, data mislabelled, ...)

Bias Systematic deviation from the true mean of the function (e.g. due to limitations of our model, ...)

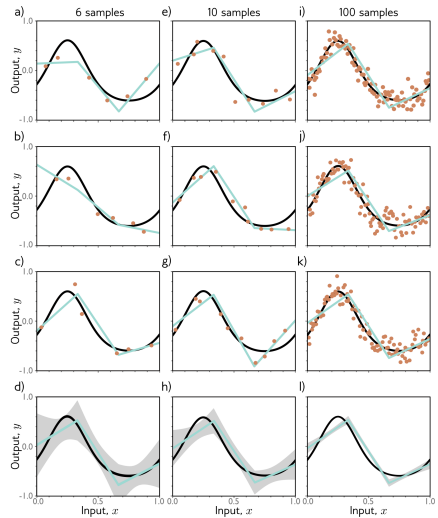
Variance Uncertainty in the fitting of our model due to limitations of the dataset (e.g. too few samples, dataset doesn't span the distribution, ...)

Evaluation: Noise, Bias and Variance

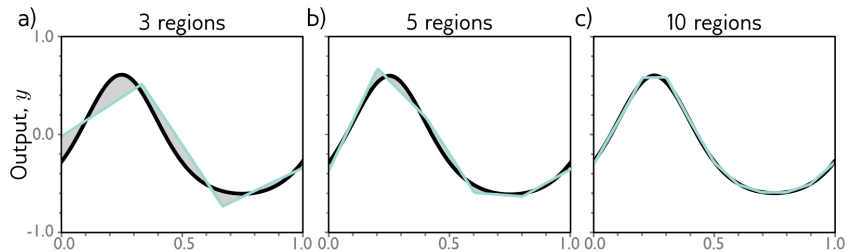
$$\mathbb{E}_{\mathcal{D}}[\mathbb{E}_y[\mathcal{L}(x, \phi)]] = \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(f(x; \phi_{\mathcal{D}}) - \overbrace{f_{\mu}(x)}^{\text{best possible model from infinite data}}\right)^2\right]}_{\text{variance}} + \underbrace{\left(f_{\mu}(x) - \overbrace{\mu(x)}^{\text{true function}}\right)^2}_{\text{bias}} + \underbrace{\sigma^2}_{\text{noise}}$$



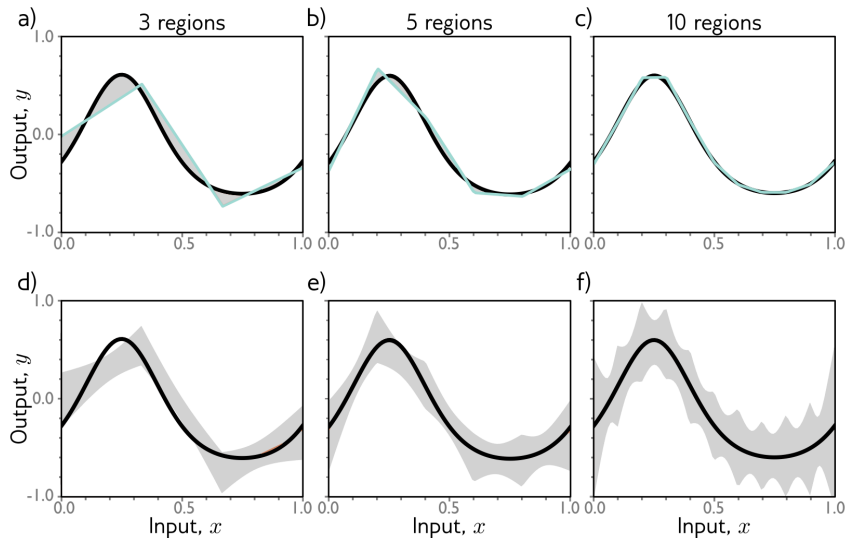
Evaluation: Variance reduction



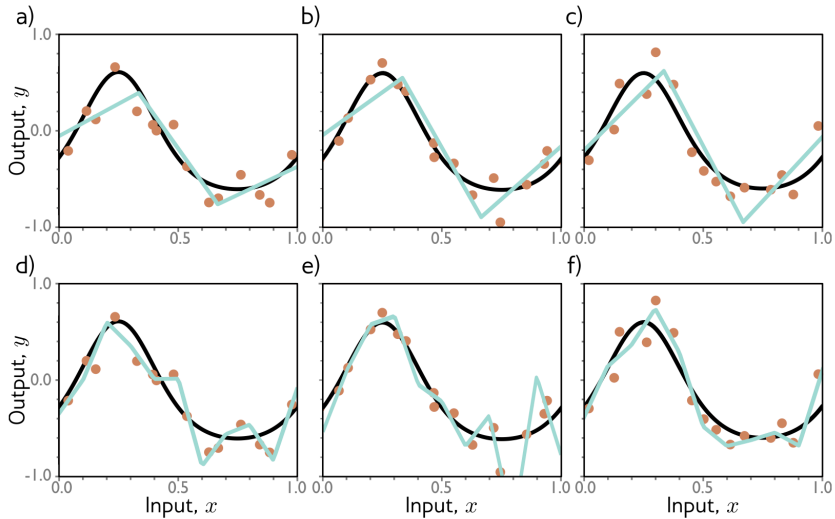
Evaluation: Bias reduction



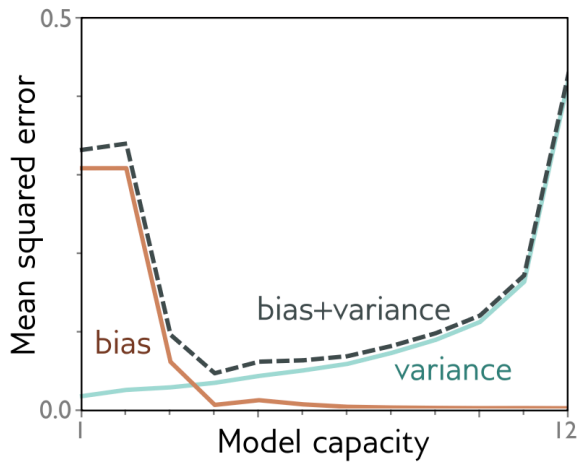
Evaluation: Bias reduction



Evaluation: Overfitting



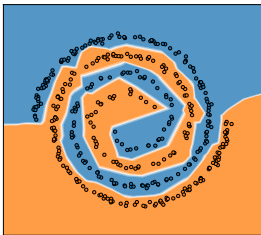
Evaluation: Bias-Variance tradeoff



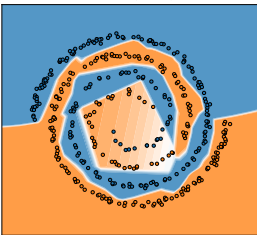
Recall: Spiral Dataset with an MLP.

Same data, varying MLP init, MLP: 32 x 2

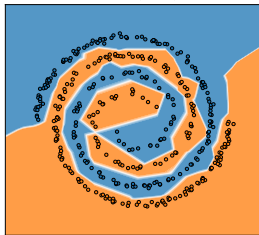
init_seed=0 (acc=1.00)



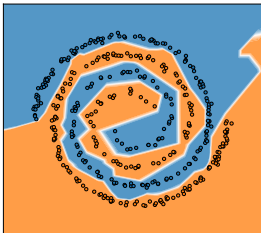
init_seed=1 (acc=0.96)



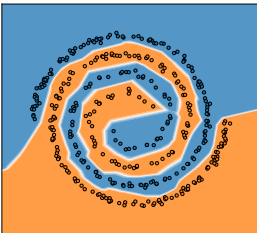
init_seed=2 (acc=1.00)



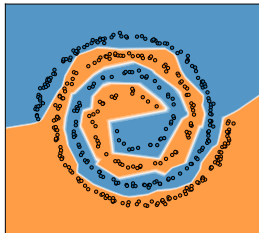
init_seed=3 (acc=1.00)



init_seed=4 (acc=1.00)



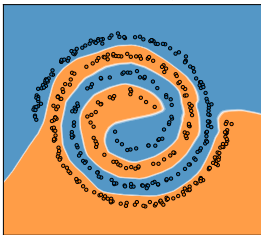
init_seed=5 (acc=1.00)



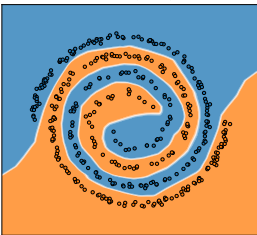
Recall: Spiral Dataset with an MLP.

Same data, varying MLP init, MLP: 64 x 4

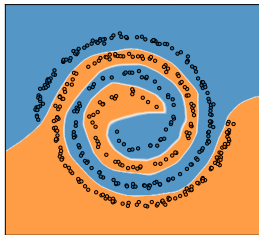
init_seed=0 (acc=1.00)



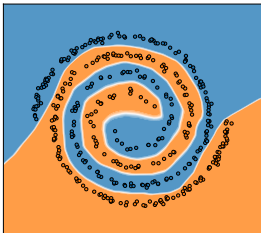
init_seed=1 (acc=1.00)



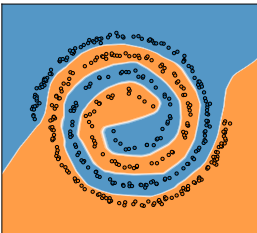
init_seed=2 (acc=1.00)



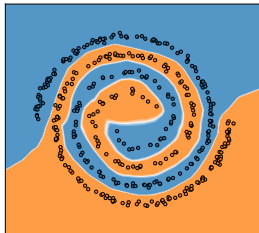
init_seed=3 (acc=1.00)



init_seed=4 (acc=1.00)



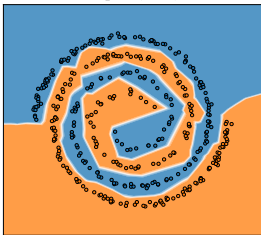
init_seed=5 (acc=1.00)



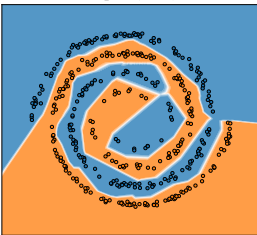
Recall: Spiral Dataset with an MLP: Different datasets..

Varying dataset, fixed MLP initialization

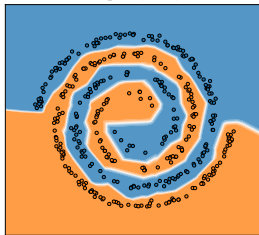
data_seed=0 (acc=1.00)



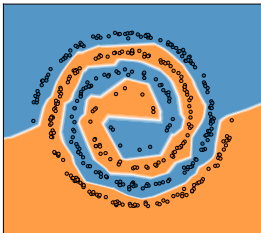
data_seed=1 (acc=1.00)



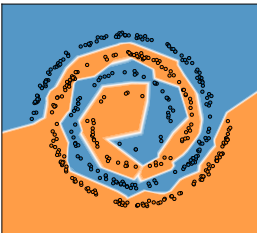
data_seed=2 (acc=1.00)



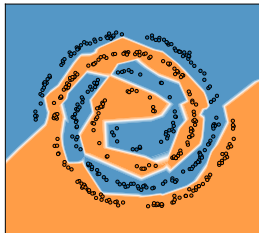
data_seed=3 (acc=1.00)



data_seed=4 (acc=1.00)



data_seed=5 (acc=1.00)



Evaluation: Statistical Significance..

- If you have some statistics training you may be familiar with the concept of **statistical significance**
- At high-level, this concerns the overlapping of (error) distributions and whether you could distinguish reliably between two distributions

Evaluation: Statistical Significance..

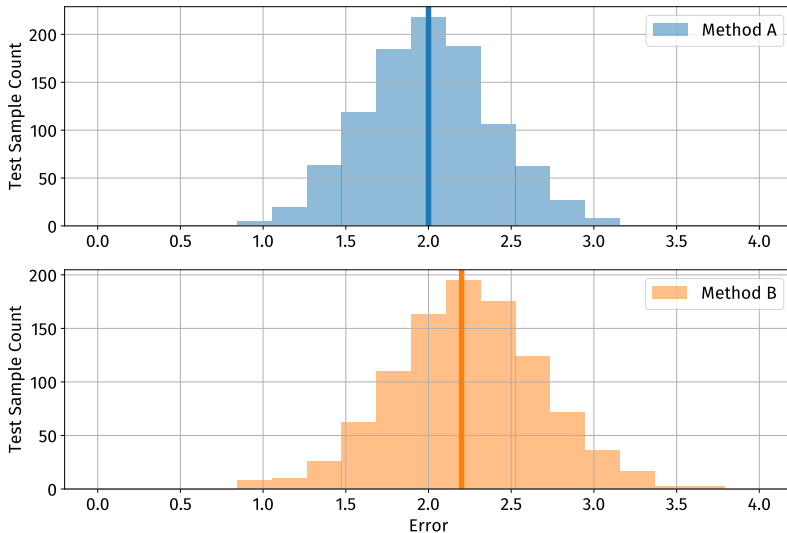
- If you have some statistics training you may be familiar with the concept of **statistical significance**
- At high-level, this concerns the overlapping of (error) distributions and whether you could distinguish reliably between two distributions
- It is possible to conduct formal tests...
- *My Advice:* try to avoid this as it is prone to all kinds of subtle decisions and arguments. **Better to show the raw data in a useful form and people can perform their own assessment...**

Evaluation: Example with statistical tests..

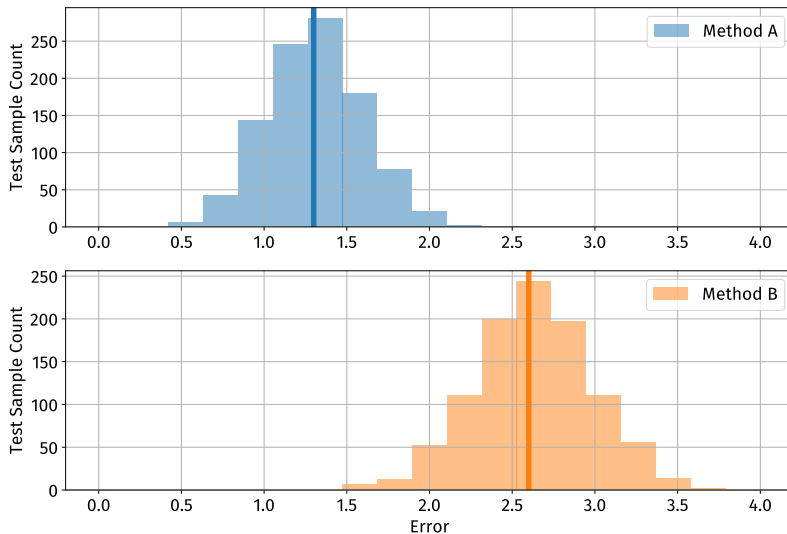
Table 1. Mean gap performance for various test functions; higher is better. The upper table shows the results after 50 objective function evaluations and the lower table after 100 evaluations. Due to computational cost, Warped GP results are only reported for 50 evaluations. Methods not significantly different from the best performing method with respect by a two-sided paired Wilcoxon signed-rank test at a 5% significance level over 20 repetitions are shown in bold (Malkomes & Garnett, 2018). For results in terms of regret, see the supplement.

Benchmark	Evals	Dim	Properties	GP	Warped GP	Homosced GP	Heterosced GP	LGP
Hartmann	50	6	boring	0.959	0.537	0.881	0.973	0.937
Griewank	50	2	oscillatory	0.914	0.493	0.752	0.913	0.897
Shubert	50	2	oscillatory	0.378	0.158	0.378	0.480	0.593
Ackley $[-10, 30]^d$	50	2	complicated, oscillatory	0.924	0.274	0.892	0.912	0.927
Cross In Tray	50	2	complicated, oscillatory	0.954	0.385	0.929	0.977	0.945
Holder table	50	2	complicated, oscillatory	0.939	0.896	0.900	0.931	0.993
Corrupted Holder Table	50	2	complicated, oscillatory	0.741	0.798	0.826	0.729	0.896

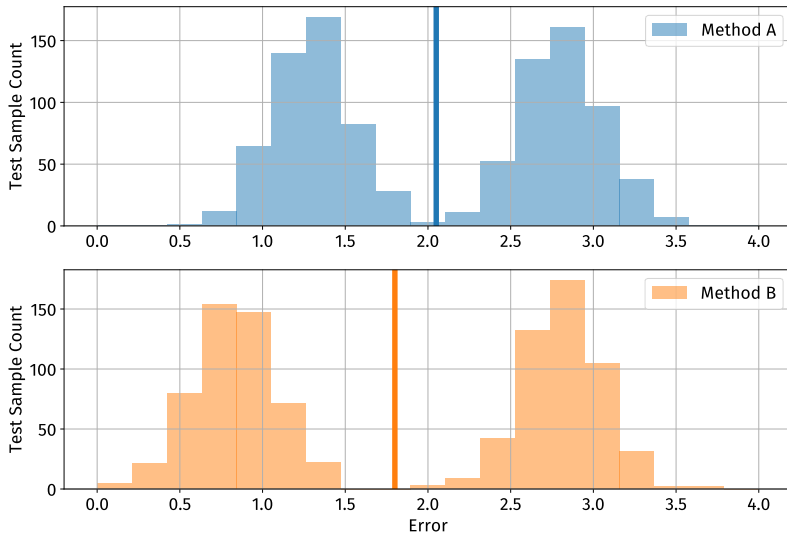
Evaluation: Look at test error distributions (e.g. histograms..)



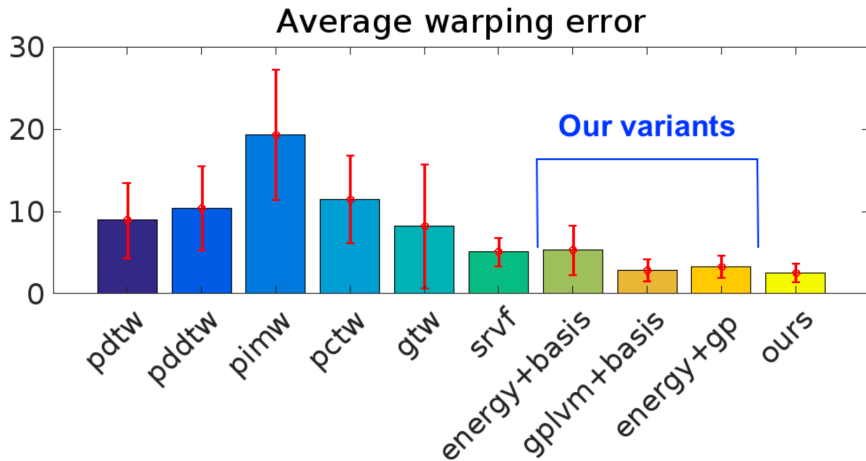
Evaluation: Look at test error distributions (e.g. histograms..)



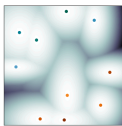
Evaluation: Look at test error distributions (e.g. histograms..)



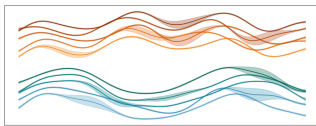
Evaluation: Alignments example with error bars..



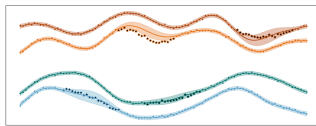
Evaluation: Alignments example with error bars..



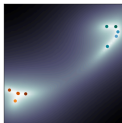
(a) MTGP $\mathbf{Z}$



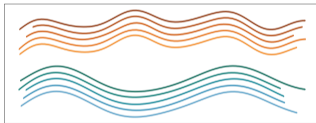
(b) MTGP function posteriors (unaligned)



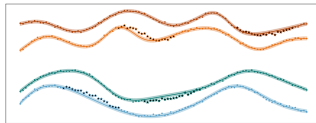
(c) MTGP missing data examples



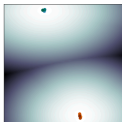
(d) GP-LVA $\mathbf{Z}$



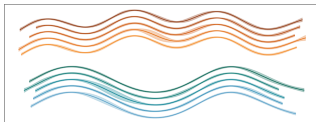
(e) GP-LVA function posteriors (aligned)



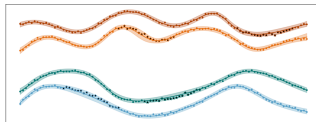
(f) GP-LVA missing data examples



(g) AMTGP $\mathbf{Z}$

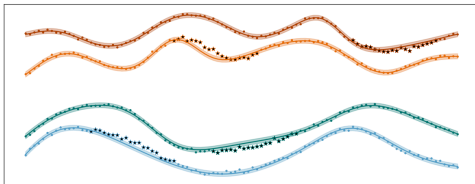


(h) AMTGP function posteriors (aligned)

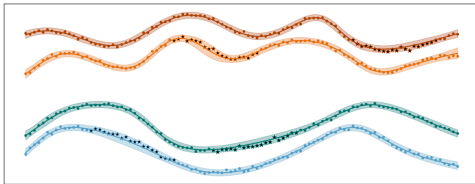


(i) AMTGP missing data examples

Evaluation: Alignments example with error bars..

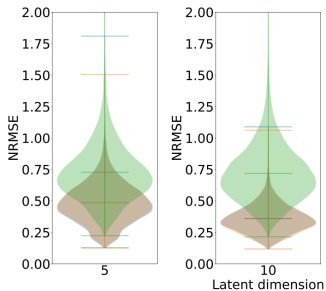


(f) GP-LVA missing data examples

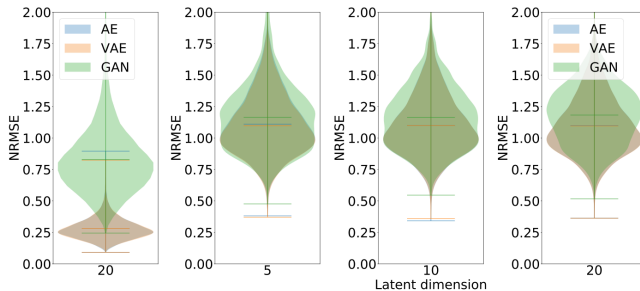


(i) AMTGP missing data examples

Evaluation: Example with histograms (e.g. violin plots)..



(a) MNIST dataset



(b) Shapes dataset

Evaluation: Histograms with illustrations!

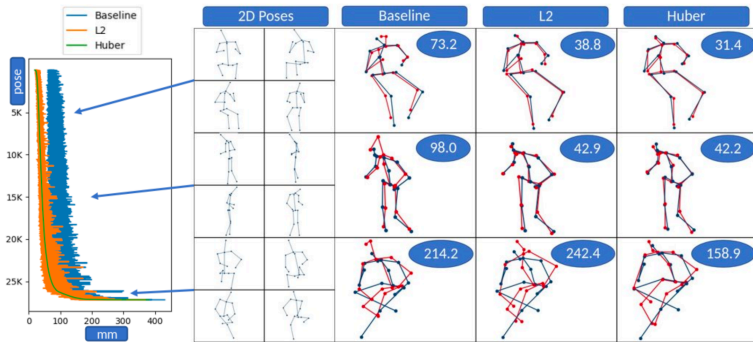


Figure 1: Error distribution on Human36M using our multi-camera model. Results are included for the untrained (Baseline) network and for the learnt shapes for the energy functions using both the ℓ_2 and the Huber loss. For visual clarity, we sort the datapoints by order of increasing error for the Huber case (i.e. our most effective approach). All models perform well on typical input instances, and fail in a limited number of cases. These outliers, however, have a noticeable, negative impact on the average error. We also provide reconstruction examples from the low, medium and tail part of the distribution.

Bayesian Machine Learning: Simple Example

Overview...

Overview

Motivation

No Free Lunch

Whirlwind Introduction to Inverse Probabilities

Model Selection

Evaluation

Bayesian Machine Learning: Simple Example

Why don't we do Model Selection in Machine Learning?

Conclusions

Switch to demo notebook..

Why don't we do Model Selection in
Machine Learning?

Overview...

Overview

Motivation

No Free Lunch

Whirlwind Introduction to Inverse Probabilities

Model Selection

Evaluation

Bayesian Machine Learning: Simple Example

Why don't we do Model Selection in Machine Learning?

Conclusions

Why don't we do Model Selection in (large-scale) ML all the time?

- History?
- Ablation Studies?
- Philosophical / Paradigm?

Why don't we do Model Selection in (large-scale) ML all the time?

- History?
- Ablation Studies?
- Philosophical / Paradigm?
- It's difficult to do!
- Integrals are often difficult or considered expensive

Why don't we do Model Selection in (large-scale) ML all the time?

- History?
- Ablation Studies?
- Philosophical / Paradigm?
- It's difficult to do!
- Integrals are often difficult or considered expensive
- Sometimes this is true, sometimes you can be clever/approximate

Why don't we do Model Selection in (large-scale) ML all the time?

- History?
- Ablation Studies?
- Philosophical / Paradigm?
- It's difficult to do!
- Integrals are often difficult or considered expensive
- Sometimes this is true, sometimes you can be clever/approximate

In era of **empirical** ML, how you evaluate is
really **important!**

Conclusions

Overview...

Overview

Motivation

No Free Lunch

Whirlwind Introduction to Inverse Probabilities

Model Selection

Evaluation

Bayesian Machine Learning: Simple Example

Why don't we do Model Selection in Machine Learning?

Conclusions

Conclusions

- Hopefully motivated the need for uncertainty in ML (?)
- These techniques are important!
- Make ML safe, trustworthy and robust for applications
- Rigour in experimental validation
- There is still much work to be done here!

Important slide acknowledgements!

Illustrations taken from the excellent new text book from Simon Prince:

Understanding Deep Learning, Simon J.D. Prince, MIT Press

Final draft available on the website:

<https://udlbook.github.io/udlbook/>

Further Reading

- Understanding Deep Learning, Simon J.D. Prince
 - <https://udlbook.github.io/udlbook/>
- Information Theory, Inference, and Learning Algorithms, David MacKay
 - <https://www.inference.org.uk/itprnn/book.html>
- Pattern Recognition and Machine Learning, Christopher M. Bishop
 - Microsoft Website with PDF
- Computer Vision: Models, Learning, and Inference, Simon J.D. Prince
 - <http://www.computervisionmodels.com/>

That's all folks..