Lecture 3: Empirical Methods for Model Selection

Attendance code: BDUJLBL6

Iain Styles

18 October 2019

Learning Outcomes

By the end of this lecture you should be able to:

- Understand the solutions to the first assignment
- Understand and be able interpret and apply the principles of empirical model selection

Assignment 0

- Not assessed
- Tested core mathematical knowledge
- Mean score 90% with the median and mode marks being 100%

Question 1

A continous probability density function has which of the following properties? Choose all that apply.

1. It must be less than 1 everywhere. (10%)
2. It must be greater than or equal to 0 everywhere. (83%)
3. The area underneath must be equal to exactly 1. (90%)
4. Its maximum value must be exactly 1. (7%)

Solution.

1. Not true. The requirements for the total area under the PDF to be equal to one does not prevent it from being greater than one.
2. True. Negative probability has no meaning in the classical framework of probability.
3. True, because only outcomes in the specified sample space can occur.
4. Not true, it is only the area that matters.

Question 2

A discrete probability distribution has which of the following properties?

1. Individual outcomes can have a probability of greater than one. (3%)
2. All outcomes must have a probability that is greater than or equal to zero. (81%)
3. The sum of the probabilities of all of the outcomes must be equal to one. (94%)
4. All outcomes must have a probability that is less than or equal to one. (84%)

Solution.

1. Not true. This would imply that an event was “more than certain”.
2. True. Negative probability have no meaning in the classical framework of probability.
3. True, because only outcomes in the specified sample space can occur.
4. True, follows from (b) and (c).

Question 3

What is the length (magnitude) of the vector $\mathbf{x} = \begin{pmatrix} 6 \\ 2 \\ 4 \end{pmatrix}$?

Give your answer to two decimal places.

Solution.

We answer this using Pythagoras' theorem: $L = \sqrt{x_1^2 + x_2^2 + x_3^2} = \sqrt{6^2 + 2^2 + 4^2} = \sqrt{36 + 4 + 16} = \sqrt{56} = 7.48.$ (94%)

Question 4

The product **XY** of the matrices $\mathbf{X} = \begin{pmatrix} 1 & 3 \\ 4 & 4 \end{pmatrix}$ and $\mathbf{Y} = \begin{pmatrix} 3 & 1 \\ 2 & 5 \end{pmatrix}$ is a 2×2 matrix. What are the values of its elements?

Solution.

Using the rules of matrix multiplication, $M_{ij} = \sum_k X_{ik} Y_{kj}$, we have

$$\mathbf{XY} = \begin{pmatrix} 1 & 3 \\ 4 & 4 \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ 2 & 5 \end{pmatrix} = \begin{pmatrix} 1 \times 3 + 3 \times 2 & 1 \times 1 + 3 \times 5 \\ 4 \times 3 + 4 \times 2 & 4 \times 1 + 4 \times 5 \end{pmatrix} \begin{pmatrix} 9 & 16 \\ 20 & 24 \end{pmatrix}$$

top-left: 95%, top-right:96%, bottom-left: 96%, bottom-left: 96%

Question 5

Matrix **X** has 13 rows and 6 columns. Matrix **Y** has 6 rows and 5 columns. Their product **XY** has ? rows and ? columns.

Solution.

Again, from the rules of matrix multiplication, $M_{ij} = \sum_k X_{ik} Y_{kj}$, $\mathbf{M} = \mathbf{XY}$ must have the same number of rows as **X** (13) and the same number of columns as **Y** (5).

Rows: 98%, Columns: 97%

Question 6

A class of students have to choose from a range of options. Two of those options are Machine Learning and Computer Graphics.

- The proportion of the class that chose Computer Graphics is 0.4
- The proportion of the class that chose Machine Learning is 0.6
- Half of those who chose Computer Graphics also chose Machine Learning.

What is the probability that a student who chose Machine Learning also chose Computer Graphics?
Give your answer accurate to two decimal places.

Solution

Denoting "Computer Graphics" as CG and "Machine Learning" as ML, we have, from the question, that $P(CG = 0.4)$, $P(ML) = 0.6$, and $P(ML|CG) = 0.5$. It then follows from Bayes theorem that $P(CG|ML) = P(ML|CG)P(CG)/P(ML) = 0.5 \times 0.4/0.6 = 0.33$
82% got this right

Question 7

What is the inverse of matrix $\mathbf{X} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$?

1. $\mathbf{X}^{-1} = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$ (0%)
2. $\mathbf{X}^{-1} = \begin{pmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{pmatrix}$ (82%)
3. $\mathbf{X}^{-1} = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{4} \end{pmatrix}$ (1%)

Solution

To solve this, we test each answer by computing $\mathbf{X}\mathbf{X}^{-1}$

1. $\mathbf{X}\mathbf{X}^{-1} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} 5 & 11 \\ 11 & 20 \end{pmatrix}$
2. $\mathbf{X}\mathbf{X}^{-1} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ ()
3. $\mathbf{X}\mathbf{X}^{-1} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{4} \end{pmatrix} = \begin{pmatrix} \frac{5}{3} & 1 \\ \frac{13}{3} & \frac{5}{2} \end{pmatrix}$

Only case (b) gives the identity matrix and this is therefore the solution.

Question 8

Two discrete variables X and Y have the joint probability distribution shown in the table below.

	$X = x1$	$X = x2$	$X = x3$
$Y = y1$	0.10	0.10	0.20
$Y = y2$	0.30	0.10	0.20

What is $P(X|Y = y2)$?

1.

$X = x1$	$X = x2$	$X = x3$
0.30	0.10	0.20

(12%)

2.

$X = x1$	$X = x2$	$X = x3$
0.25	0.25	0.50

(2%)

3.

$X = x1$	$X = x2$	$X = x3$
0.50	0.17	0.33

(84%)

Note that the values in the tables above are accurate to two decimal places.

Solution

We need the product rule of probability here:

$P(X, Y) = P(X|Y)P(Y)$. From this, we have

Question 9

A function $f(x)$, defined for $-\infty \leq x \leq \infty$, has its minimum value when $x = 0$. Which one of the following statements is true?

1. The gradient of $f(x)$ at $x = 0$ is equal to the value of $f(x)$ at $x = 0$. (93%)
2. The value of $f(x)$ is zero at $x = 0$. (2%)
3. The gradient of $f(x)$ is zero at $x = 0$. (5%)

Solution

The only way that a function can take a minimum value away from the edges of its domain is for there to be a turning point where the gradient is zero (c). The other choices do not imply a minimum. Although it was not explicitly stated in the question, there are some special cases where this may not be true, and there are stronger conditions needed on the function. For example, $f(x) = 1/x$ has a singularity at $x = 0$ and the minimum values tends to $-\infty$. The question should have stated that the function is required to be continous and differentiable (i.e. smooth).

Question 10

A column vector **v** has components v_i. The magnitude, or length of v is given by the formula $L = \sqrt{\sum_i v_i^2)}$. This can also be written as:

1. $L = \mathbf{v}\mathbf{v}^{\mathrm{T}}$ (89%)
2. $L = \sqrt{\mathbf{v}\mathbf{v}^{\mathrm{T}}}$ (3%)
3. $L = \sqrt{\mathbf{v}^{\mathrm{T}}\mathbf{v}}$ (6%)
4. $L = \mathbf{v}^{\mathrm{T}}\mathbf{v}$ (1%)

Solution

Noting again the rules of matrix multiplication and treating a column vector as an $N \times 1$ matrix, it follows that $\sum_i v_i^2 = \sum_i v_{1i} v_{i1} = \mathbf{v}^{\mathrm{T}}\mathbf{v}$. Taking the square root leads to the correct answer (c).

Selecting and Evaluating Models

- We try to find function $f(\mathbf{x})$ to match underlying data generating function $h(x)$
- If we know h we can easily choose f
- But normally we do not know h
- f then has to be determined experimentally
- *Validation and Testing*
- Checking the model's ability to predict unseen data
- Two main approaches

Train–Validate–Test

- Common for very large data sets
- Often used in machine learning competitions
- Some of the data used to train the model
- Some of the data used to evaluate whether that model is any good
- Some of the data used to test what we think is the best model
- Especially popular with large volume of data

Train–Validate–Test

- Partition a dataset $\mathcal{D}$ into:
  - A training set $\mathcal{T}$, randomly sampled from $\mathcal{D}$.
  - An validation set $\mathcal{V}$, randomly sampled from $\mathcal{D} - \mathcal{T}$.
  - A test, or evaluation set $\mathcal{E} = \mathcal{D} - \mathcal{T} - \mathcal{V}$.

Train–Validate–Test

- Partition a dataset $\mathcal{D}$ into:
  - A training set $\mathcal{T}$, randomly sampled from $\mathcal{D}$.
  - An validation set $\mathcal{V}$, randomly sampled from $\mathcal{D} - \mathcal{T}$.
  - A test, or evaluation set $\mathcal{E} = \mathcal{D} - \mathcal{T} - \mathcal{V}$.
- Define a set of models $\{\mathcal{M}_i\}_{k=1}^{K}$
- And a loss function $\mathcal{L}$ (least-square for regression)

The Goal

- Training set $\mathcal{T}$ used to optimise model parameters
- Validation set $\mathcal{V}$ used to select optimal model – *hyperparameter optimisation*
- Select the choice of hyperparameters that best allows the model learned on $\mathcal{T}$ to generalise to $\mathcal{V}$
- Evaluate on $\mathcal{E}$ to assess how well the model performs on unseen data
- Ultimate test of its ability to generalise
- Guards against overfitting of the hyperparameters to $\mathcal{V}$

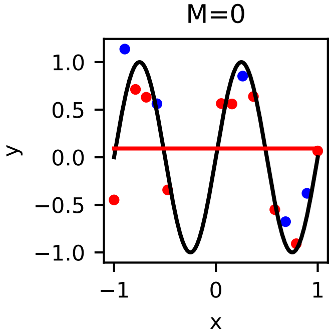
The Algorithm

```
Data: Set of models $\{\mathcal{M}_i\}_i$
Data: Dataset $\mathcal{D}$ split into training ($\mathcal{T}$), validation ($\mathcal{V}$), and
      test/evaluation ($\mathcal{E}$) sets.
Result: Identification of model $\mathcal{M}*$ with best predictive power.
for each model $\mathcal{M}_i$ do
    Train $\mathcal{M}_i$ on $\mathcal{T}$;
    Compute model loss $\mathcal{L}_{\mathcal{T}}$ on training set $\mathcal{T}$;
    Compute model loss $\mathcal{L}_{\mathcal{V}}$ on evaluation set $\mathcal{V}$;
end
Select model $\mathcal{M}*$ with best overall performance on training and
validation sets.;
Compute loss on test set $\mathcal{E}$ to determine final model performance;
```

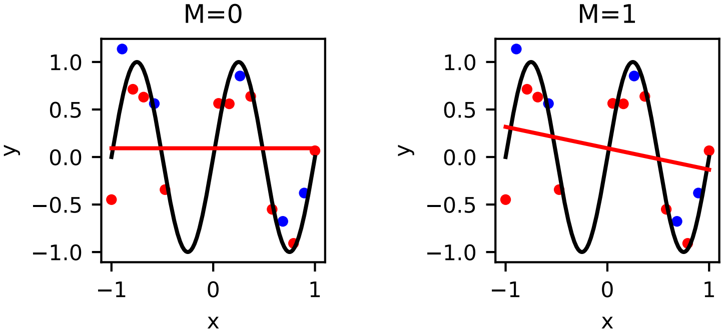
Notes

- $\mathcal{T}$ must be big enough to fully represent the data: an 80-10-10 split is quite common
- $\mathcal{D}$ must be randomised to ensure $\mathcal{T}$, $\mathcal{V}$, $\mathcal{E}$ represent the data
- Eg: sample from across the data's domain x and range y
- Example: twenty data points and split them into a training set $\mathcal{T}$ of ten points, a validation set $\mathcal{V}$ of five points, and a test set $\mathcal{E}$ of 5 points.

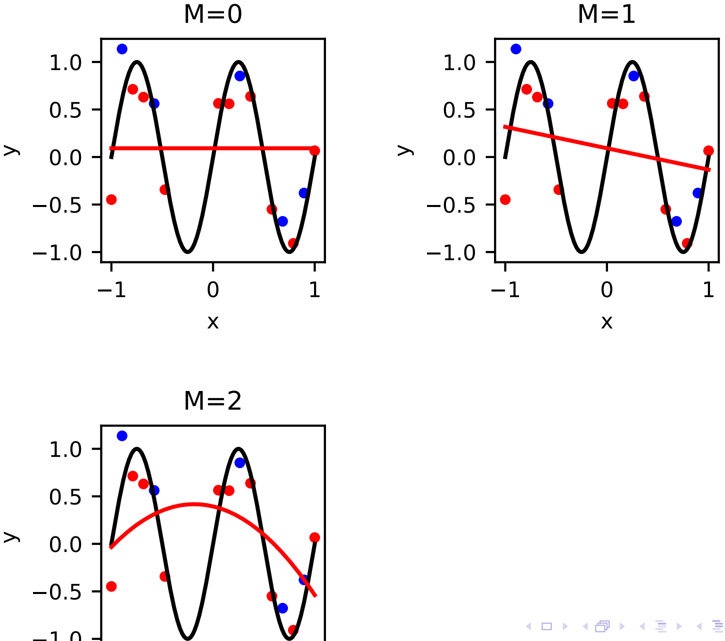
Visualisation of Training



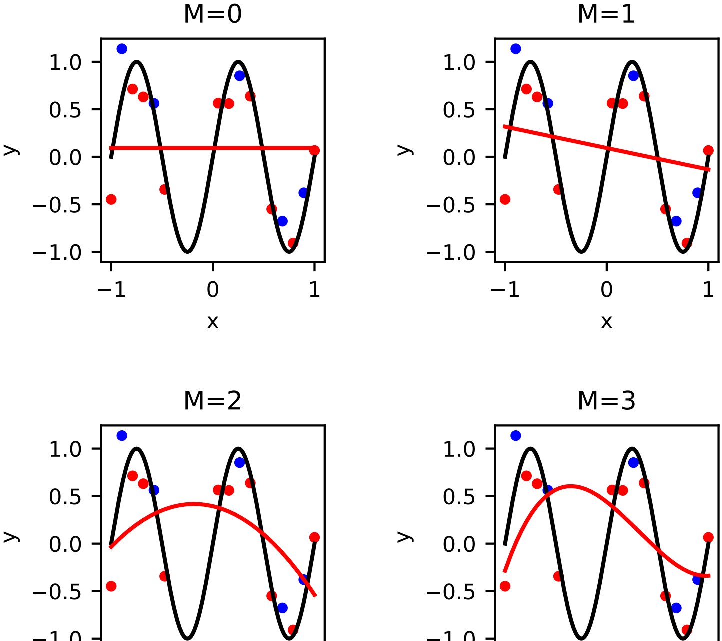
Visualisation of Training



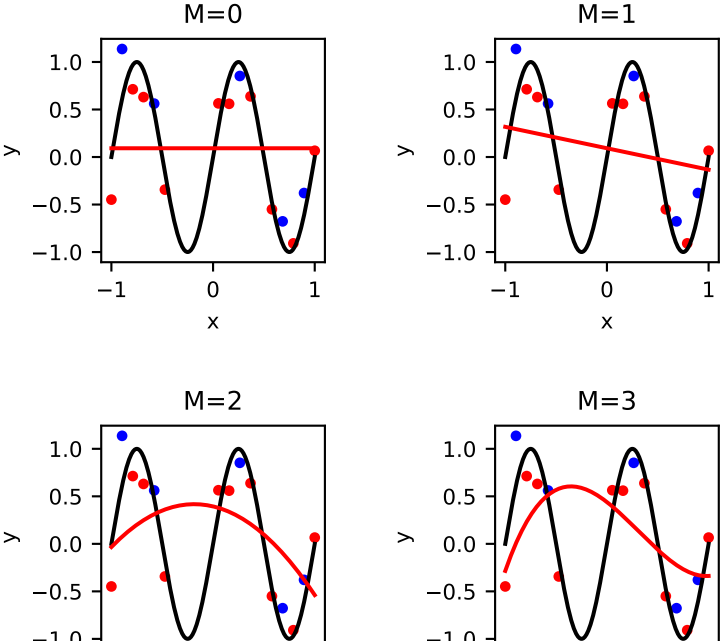
Visualisation of Training



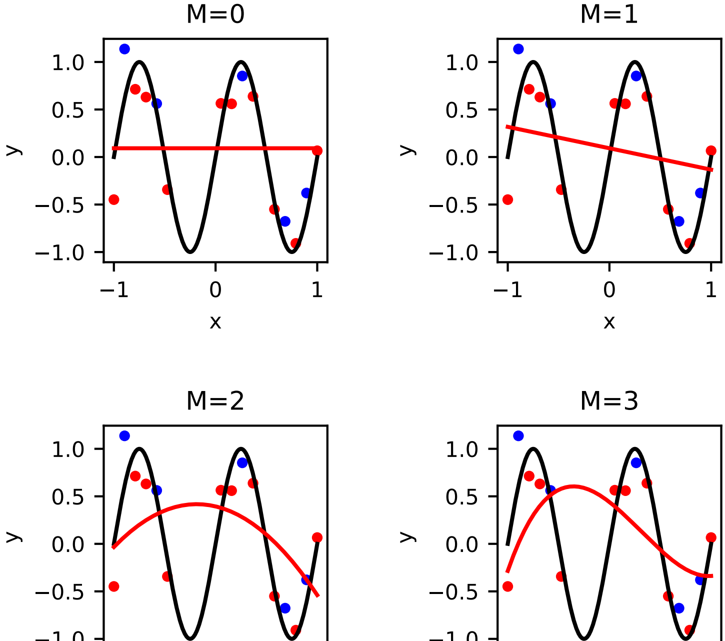
Visualisation of Training



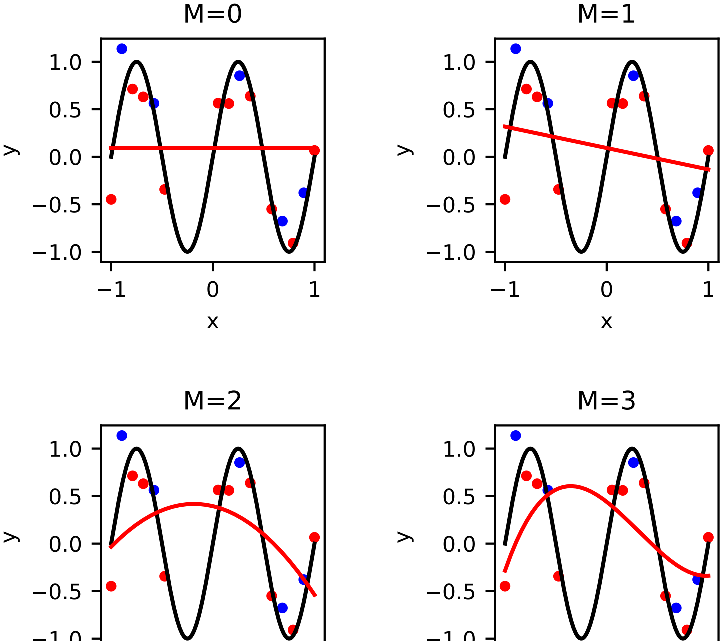
Visualisation of Training



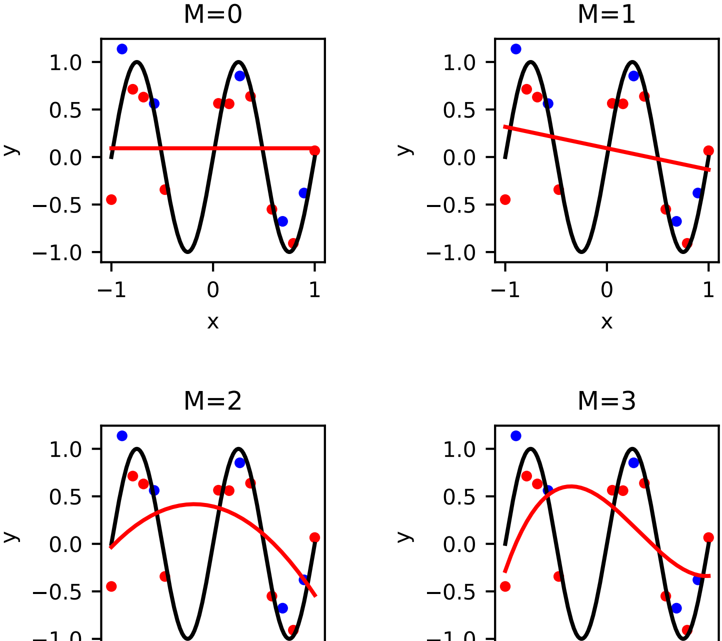
Visualisation of Training



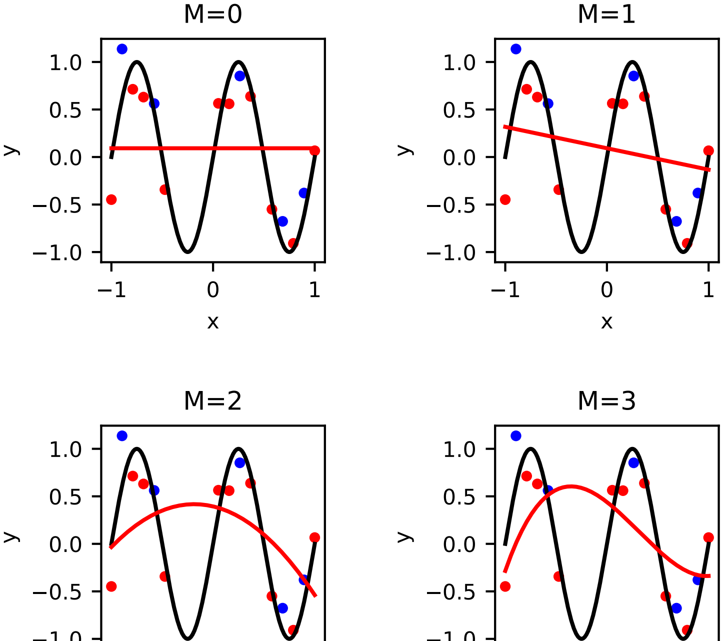
Visualisation of Training



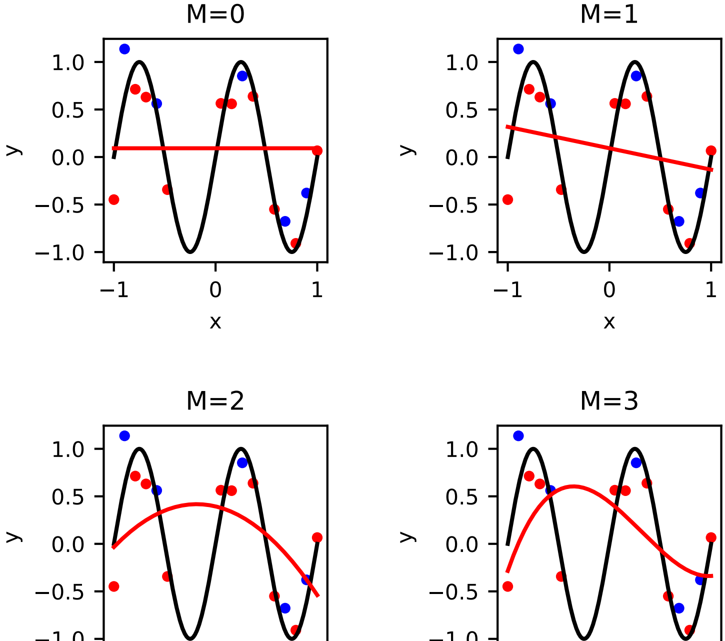
Visualisation of Training



Visualisation of Training

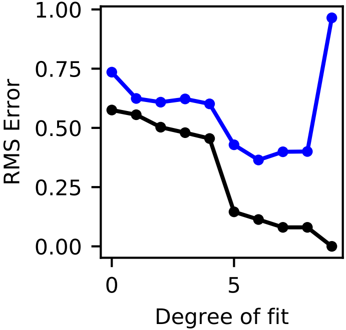


Visualisation of Training



Evaluation

- Training error continues to improve
- Validation error also improves but then gets dramatically worse
- Model is overfitting the training data
- Models trend + noise so cannot generalise
- Occam's razor: choose simplest model that performs well
- $M = 5$?



Notes

- Validation set is used to select the best model
- Test set is used right at the end to ensure that the model generalises beyond the validation set
- Avoids hyperparameter overfitting
- Test set error here is 0.68 – comparable to validation set.
- What if we had even less data?

Cross validation

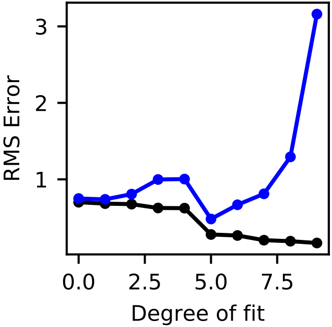
- Small data – may not be able to adequately split into representative groups
- In *cross-validation* we split the data into K "folds".
- Train models on $K - 1$ of the folds
- Validate on the remaining fold
- Use each of the folds in turn as the validation set
- Select the model that gives the best average performance

The Cross Validation Algorithm

```
Data: Set of models $\{\mathcal{M}_i\}_i$
Data: Dataset $\mathcal{D}$ split into cross-validation ($\mathcal{V}$), and
      test/evaluation ($\mathcal{E}$) sets.
Data: Number of folds, $K$
Result: Identification of model $\mathcal{M}*$ with best predictive power.
Divide $\mathcal{C}$ into $K$ folds $\{c_k\}_{k=1}^{K}$ such that $\mathcal{C} = \bigcup_{k=1}^{K} c_k$;
for each model $\mathcal{M}_i$ do
    for $k = 1 \to K$ do
        Train $\mathcal{M}_i$ on training set $\mathcal{C} - c_k$;
        Compute model loss $\mathcal{L}_{\mathcal{T}}$ on training set $\mathcal{C} - c_k$;
        Compute model loss $\mathcal{L}_{\mathcal{V}}$ on evaluation fold $c_k$;
    end
end
Select model $\mathcal{M}*$ with best overall performance on training and
validation sets;
Compute loss on test set $\mathcal{E}$ to determine final model performance;
```

Evaluation

- K=5-fold cross validation on 20-pt dataset
- Training set is larger – 16 points in each round
- Results are averaged over the folds
- Validation error also improves but then gets dramatically worse
- Similar conclusions can be drawn: models of order 5 perform well on both the training and validation sets



Summary

- Splitting the data up allows us to experimentally determine the optimal model
- Can be computationally expensive - especially cross validation
- Next time – controlling models with prior knowledge
- Regularisation, leading into a Bayesian approach to regression