Machine Learning 1.08: Is it working?

Tom S. F. Haines
T.S.F.Haines@bath.ac.uk

Failure

How can your machine learning system fail?

Failure

How can your machine learning system fail?

- Underfitting
- Overfitting
- Bad data
- Incorrect use

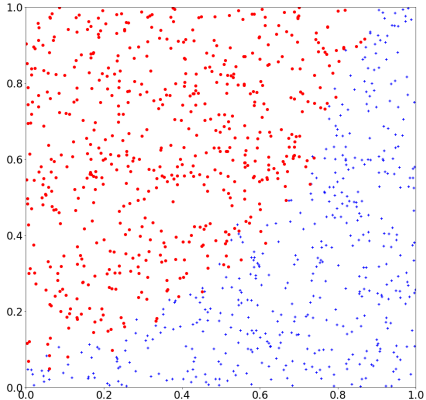
Failure

How can your machine learning system fail?

- Underfitting
- Overfitting
- Bad data
- Incorrect use

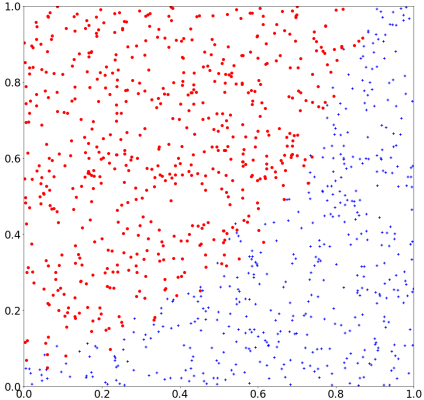
How do you know?

Underfitting & Overfitting

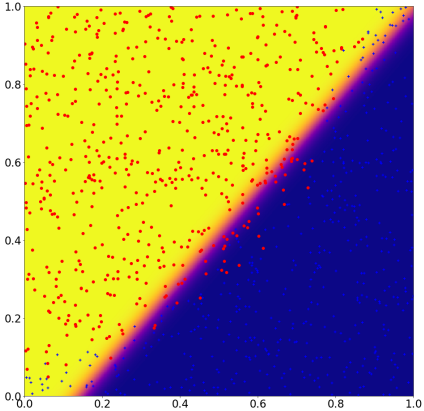


- Curved.
- Classes overlap.
- How would you divide them?

Underfitting & Overfitting

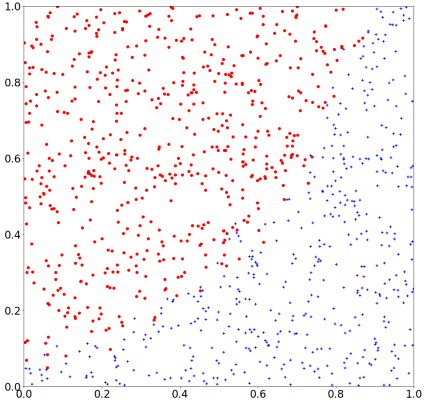


- Curved.
- Classes overlap.
- How would you divide them?

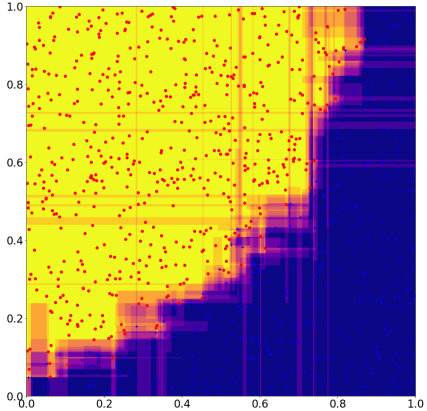


- Underfitting

Underfitting & Overfitting

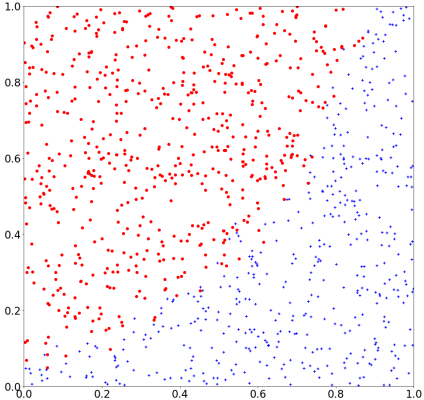


- Curved.
- Classes overlap.
- How would you divide them?

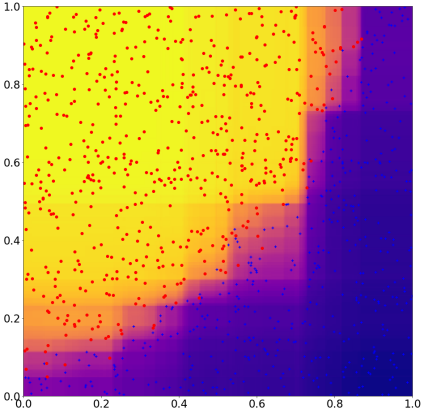


- Overfitting

Underfitting & Overfitting

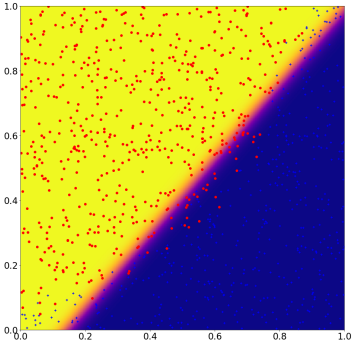


- Curved.
- Classes overlap.
- How would you divide them?

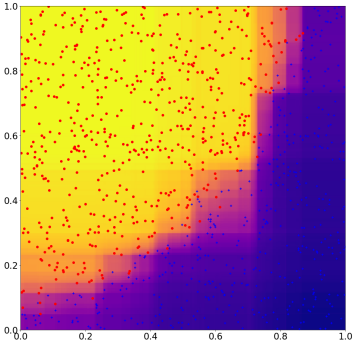


- Balanced

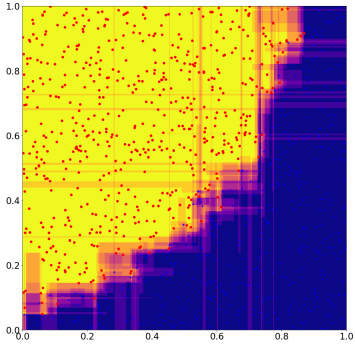
Underfitting & Overfitting



- Underfitting
- Logistic regression



- Balanced
- Tuned random forest.
- (scikit learn, `min_impurity_decrease=0.008`, `n_estimators=512`)

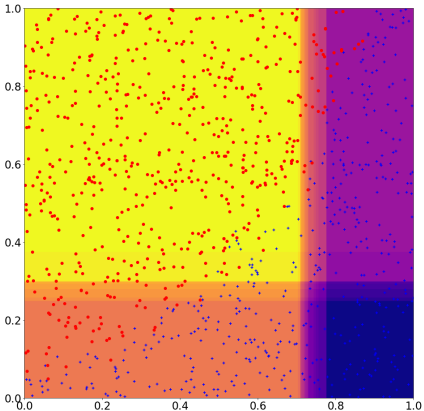


- Overfitting
- Badly tuned random forest.
- (scikit learn, default parameters)

Underfitting Causes

- Weak model
- Bad fitting (left, random forest again)

- Bad data
- Insufficient data



Overfitting Causes

- Powerful model – capable of modelling the noise.
  +
- Insufficient **regularisation**.
  Regularisation $\sim$ smoothing out the noise.
  (Subject of next lecture!)

UNIVERSITY OF BATH

Overfitting Causes

- Powerful model – capable of modelling the noise.
  +
- Insufficient **regularisation**.
  Regularisation $\sim$ smoothing out the noise.
  (Subject of next lecture!)

- Simple version: Incorrect hyper-parameters.
  Hyper-parameters = parameters that affect algorithm behaviour.
- How to detect?

Train & Test set

- Model can't overfit on data it doesn't have!
  $\therefore$
- Split the data:
  - A **train** set, to fit the model.
  - A **test** set, to verify performance.

Train & Test set

- Model can't overfit on data it doesn't have!
  $\therefore$
- Split the data:
  - A **train** set, to fit the model.
  - A **test** set, to verify performance.
- Large gap between train/test accuracy indicates overfitting (usually).

	Accuracy	
Random Forest	Train	Test
Underfitting	79.2%	79.2%
Balanced	97.6%	95.0%
Overfitting	99.6%	94.7%

UNIVERSITY OF BATH

Hyperparameters

- Parameters = optimised to fit data.
- Hyperparameters = set by you before parameter optimisation.
  (more precise meaning for Bayesian models)

- You can of course tune the hyperparameters.
  (manually or by algorithm)

UNIVERSITY OF BATH

Hyperparameters

- Parameters = optimised to fit data.
- Hyperparameters = set by you before parameter optimisation.
  (more precise meaning for Bayesian models)

- You can of course tune the hyperparameters.
  (manually or by algorithm)

- **Do not use the test set!**
  (this mistake can be found in countless research papers)

Hyperparameters

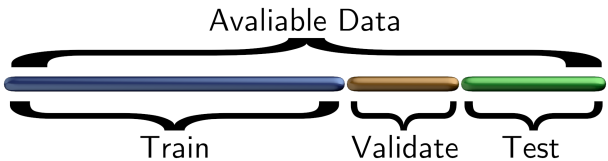
- Parameters = optimised to fit data.
- Hyperparameters = set by you before parameter optimisation.
  (more precise meaning for Bayesian models)

- You can of course tune the hyperparameters.
  (manually or by algorithm)

- **Do not use the test set!**
  (this mistake can be found in countless research papers)

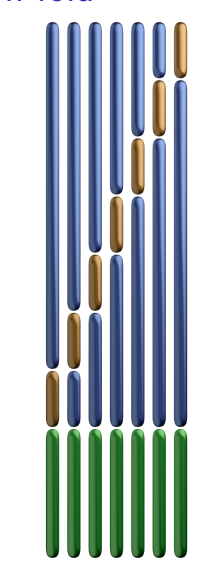
- Introduce a third set: **validation** set.
  - **train** – Give to algorithm.
  - **validation** – Objective of hyperparameter optimisation.
  - **test** – Use to report final performance.

Measuring Performance



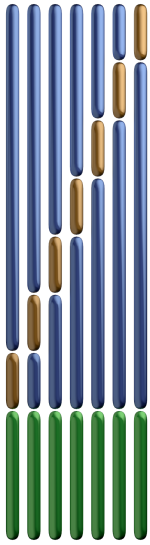
- How do we decide on split percentages?
  - Train large $\rightarrow$ Algorithm performs well.
  - Validation large $\rightarrow$ Hyperparameter optimisation gets accurate estimate, and performs well.
  - Test large $\rightarrow$ Accurate performance estimate.

- Validation and test used to take **measurements**.
- Can average multiple estimates together!



UNIVERSITY OF BATH

n-fold

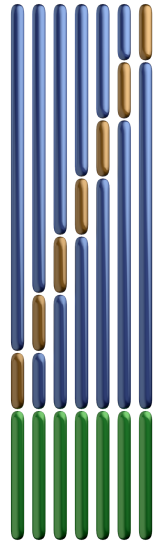


- Validation and test used to take **measurements**.
- Can average multiple estimates together!

- e.g. divide train/validation into 7-fold, use one part for validation, rest for training, and train 7 times, validation performance is average performance for all runs.
- Extend to test: Divide whole data set into n parts, run every single combination of 1 part validation, 1 part test, everything else training.

UNIVERSITY OF BATH

n-fold



- Validation and test used to take **measurements**.
- Can average multiple estimates together!

- e.g. divide train/validation into 7-fold, use one part for validation, rest for training, and train 7 times, validation performance is average performance for all runs.
- Extend to test: Divide whole data set into *n* parts, run every single combination of 1 part validation, 1 part test, everything else training.

- n-fold = n $\times$ the computation!
- Most extreme: Jackknife resampling. One exemplar left out of training; horrifically slow.

UNIVERSITY OF BATH

Out of Bag Error

- Out-of-bag error is provided by random forests, among others.
  1. Each exemplar gets estimate from all trees that didn't train on it.
  2. Tree predictions merged for each exemplar.
  3. Accuracy measured.

UNIVERSITY OF BATH

Out of Bag Error

- Out-of-bag error is provided by random forests, among others.
  1. Each exemplar gets estimate from all trees that didn't train on it.
  2. Tree predictions merged for each exemplar.
  3. Accuracy measured.

- This isn't correct – somewhere between train and test.
- Overconfident – do not trust.
  (but free to calculate, so no harm in looking)

UNIVERSITY OF BATH

Final Model

- May train algorithm thousands of times!
- Choice of n is a trade-off between accuracy / time.
- Fast computer/cluster/distributed computation really help!

- Final model: Train on entire data set.

UNIVERSITY OF BATH

Confusion Matrices

- Classification only.
- Random forest on breast cancer:

		Actual	
		False	True
Predicted	False	49	6
	True	14	159

Confusion Matrices

- Classification only.
- Random forest on breast cancer:

		Actual	
		False	True
Predicted	False	49	6
	True	14	159

- On diagonal means correct, off means false.
- Can see which classes are confused.
- An empty row is a problem.
- May want to colour code cells as a heat map!

UNIVERSITY OF BATH

Naming the Numbers

		Actual	
		False	True
Predicted	False	True Negative (TN)	False Negative (FN)
	True	False Positive (FP)	True Positive (TP)

UNIVERSITY OF BATH

Naming more Numbers

Loads of terms are used (ignore most of them):

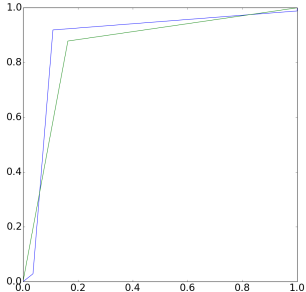
$\frac{TP}{TP+FN}$	sensitivity, **recall**, hit rate, **true positive rate**
$\frac{TN}{TN+FP}$	specificity, **true negative rate**
$\frac{TP}{TP+FP}$	**precision**, positive predictive value
$\frac{TP+TN}{TP+TN+FP+FN}$	**accuracy**
$\frac{2\times TP}{2\times TP+FP+FN}$	**F1 score**

(many more. . .)

UNIVERSITY OF BATH

ROC Curve

- Previous all assume mistakes are equally bad. Usually not!
- Receiver operating characteristic ROC) curve:



- Threshold sweep. Lets you see the tradeoff – want to be as close to the top left as possible.
- True positive rate (x-axis) against false positive rate (y-axis).
- Blue = random forest; Green = linear regression.

Performance

- What you're trying to ultimately optimise!
- A problem specific function of the confusion matrix (for classification).

- Depending on problem might be better to think in terms of:
  - Cost
  - Gain
  - Error
  - Risk

UNIVERSITY OF BATH

Group Exercise

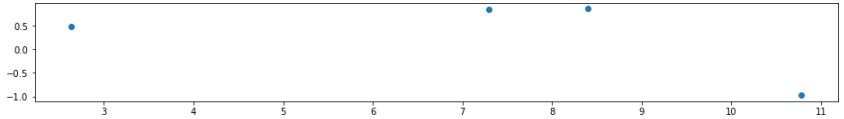
In small groups discuss how you would measure performance for:

Deciding if a bank's customer should be issued a mortgage or not.	Selecting adverts to show on a website.	Adjusting the route of a delivery driver to factor in predicted traffic conditions.
Identifying the speed limit for a self-driving car. What about detecting pedestrians? What if you could detect their age?	Predicting the probability of reoffending during sentencing, which is then factored into the prison sentence and parole conditions.	Retinopathy of Prematurity is when a baby is born before the blood vessels in their eye have fully developed. It's hard to detect and surgery is dangerous, but without surgery they will be blind.

UNIVERSITY OF BATH

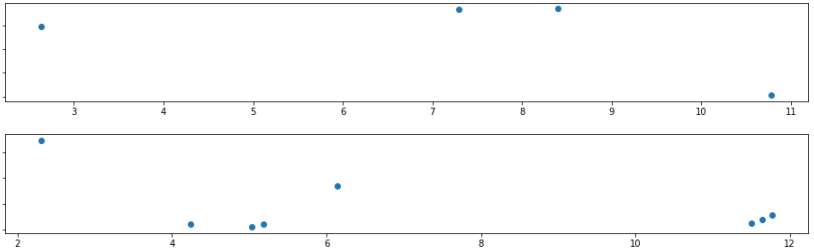
Bad Data

Bad Data – Insufficient



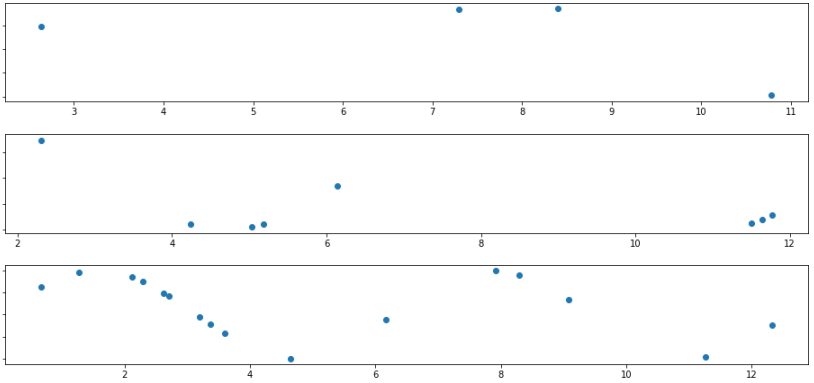
UNIVERSITY OF BATH

Bad Data – Insufficient



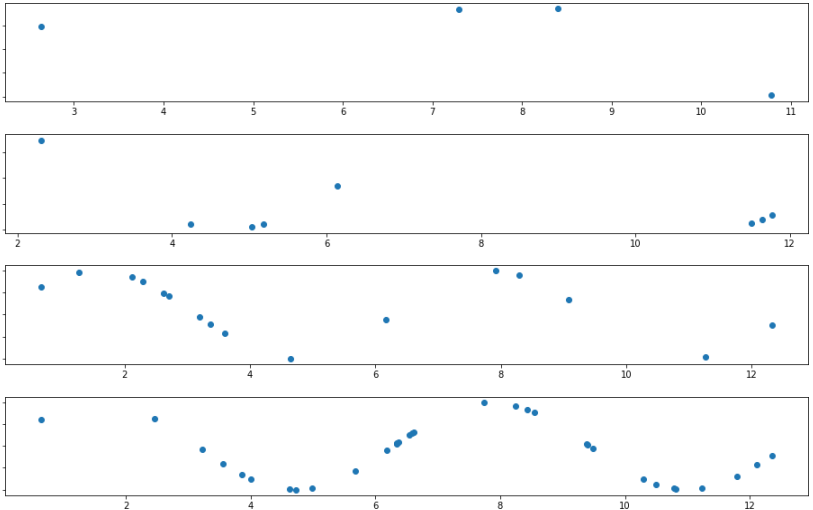
UNIVERSITY OF BATH

Bad Data – Insufficient



UNIVERSITY OF BATH

Bad Data – Insufficient



UNIVERSITY OF BATH

Bad Data – Spurious Correlation

- In 1964 a researcher was spotting M-48 tanks in images.
- Got a near perfect score.

UNIVERSITY OF BATH

Bad Data – Spurious Correlation

- In 1964 a researcher was spotting M-48 tanks in images.
- Got a near perfect score.

- Problem:
  - Tank photos were taken on cloudy day.
  - Not-tank photos were taken on a sunny day.
  - . . . so it was checking the sky colour.
- Original paper (probably!): `https://dl.acm.org/citation.cfm?doid=800257.808903`

UNIVERSITY OF BATH

Bad Data – Unbalanced

- When you train with 1000 examples of one class and 10 of another.
- Classifier can get 99% by allways predicting the larger class...
- ...and often does.

- Good example of this: `https://arxiv.org/pdf/1606.08390.pdf`

Bad Data – Selection Bias

- During WW2 the US military wanted to selective add armour to their bombers.
- Intial idea: Add it where the holes on the returning bombers were.

UNIVERSITY OF BATH

Bad Data – Selection Bias

- During WW2 the US military wanted to selective add armour to their bombers.
- Intial idea: Add it where the holes on the returning bombers were.

- Abraham Wald (a statician) pointed out that you want to put extra armour where there are no holes!
- The holes tell you where a plane can be hit and fly home.

UNIVERSITY OF BATH

Bad Data – Selection Bias

- During WW2 the US military wanted to selective add armour to their bombers.
- Intial idea: Add it where the holes on the returning bombers were.

- Abraham Wald (a statician) pointed out that you want to put extra armour where there are no holes!
- The holes tell you where a plane can be hit and fly home.

- Ever hear the claim music used to be better?
- Nice summary paper: `https://people.ucsc.edu/~msmangel/Wald.pdf`
- Original document: `http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA091073`

Bad Data – Biased Data I

- Hospital wants to use machine learning to decide if they should admit or send home patients with pneumonia.
- Train on survival rate.
- It recomends sending patients with asthma home. . . which would amost certainly kill them.

Bad Data – Biased Data I

- Hospital wants to use machine learning to decide if they should admit or send home patients with pneumonia.
- Train on survival rate.
- It recomends sending patients with asthma home. . . which would amost certainly kill them.

- Hospital policy: Any asthma patient with pneumonia is sent straight to the ICU.
- They do such a good job their survival rate is higher than those sent home!

- Study in which above problem was identified:
  `http://people.dbmi.columbia.edu/noemie/papers/15kdd.pdf`

Bad Data – Biased Data II

- Florida decides sentencing based on predicted reoffending rate.
- Trained machine learning system on past judge behaviour.

UNIVERSITY OF BATH

Bad Data – Biased Data II

- Florida decides sentencing based on predicted reoffending rate.
- Trained machine learning system on past judge behaviour.

- Turns out the judges are racist.
- If you train on human behaviour you will learn the flaws.

- Replicating racist judges: `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`
- Another example: Microsoft's racist chat bot: `https://www.theguardian.com/technology/2016/mar/26/microsoft-deeply-sorry-for-offensive-tweets-by-ai-chatbot`

Detecting Bad Data

- Visualise.
  That's it!
- But be careful – a bad visualisation can be misleading.

UNIVERSITY OF BATH

End.

- Overfitting/underfitting
- Train/test/verification
- Measuring success
- Bad data