Machine Learning 1.04: 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
- Misinterpretation
- Bad data

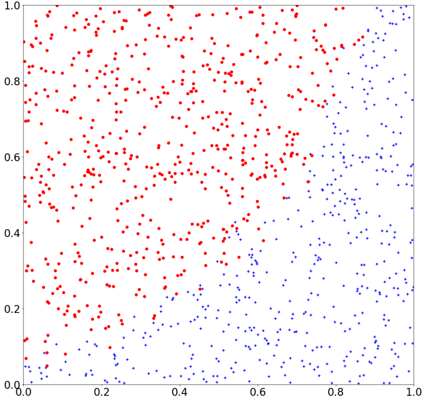
Failure

How can your machine learning system fail?

- Underfitting
- Overfitting
- Misinterpretation
- Bad data

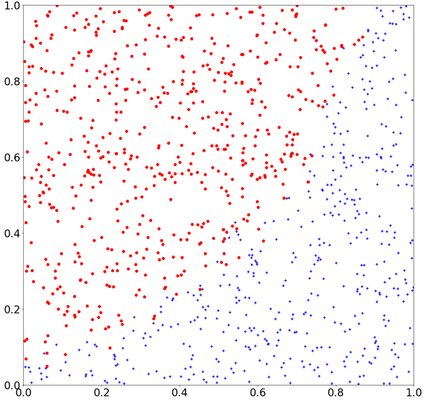
How do you know?

Underfitting & overfitting

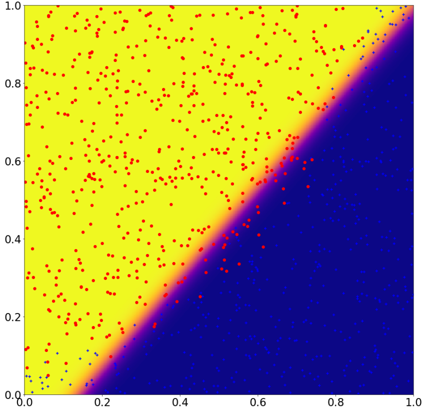


- Curved
- Classes overlap
- How would you divide them?

Underfitting & overfitting

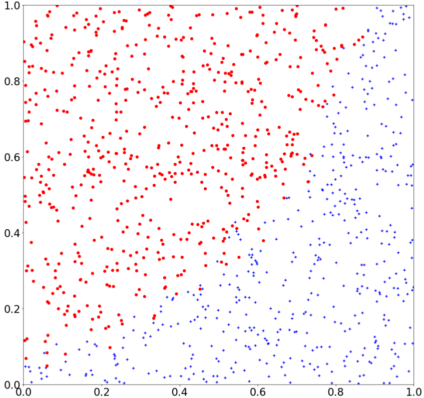


- Curved
- Classes overlap
- How would you divide them?

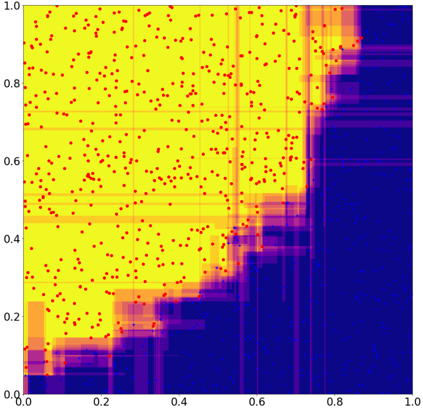


- Underfitting

Underfitting & overfitting

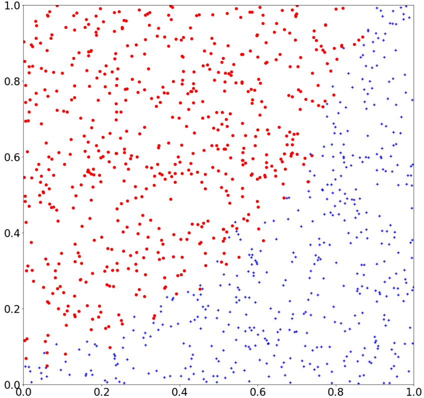


- Curved
- Classes overlap
- How would you divide them?

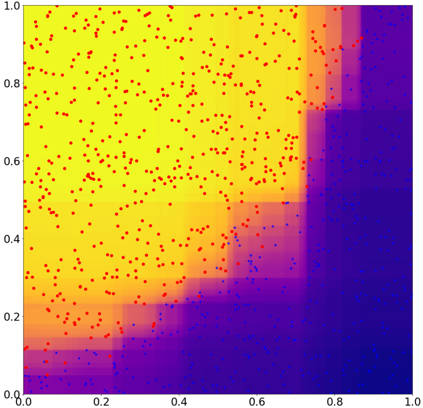


- Overfitting

Underfitting & overfitting

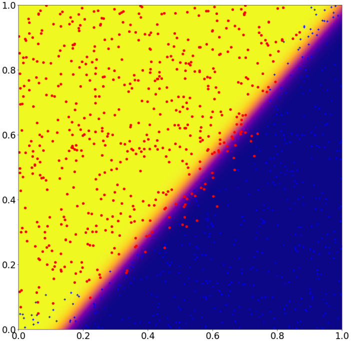


- Curved
- Classes overlap
- How would you divide them?

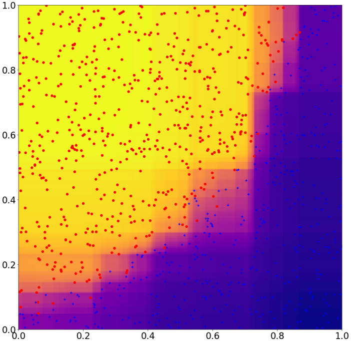


- Balanced

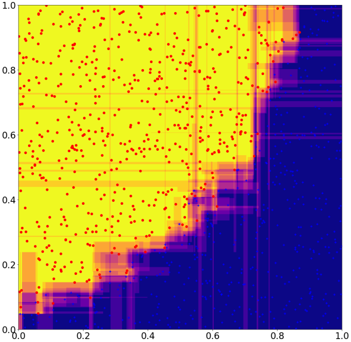
Underfitting & overfitting



- Underfitting
- Logistic regression



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

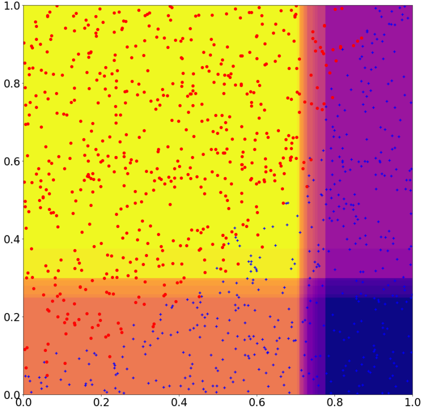


- Overfitting
- Badly tuned random forest
- (scikit learn, default parameters)

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 later lecture)

Overfitting causes

- Powerful model – capable of modelling the noise
  $+$
- Insufficient **regularisation**
  Regularisation $\sim$ smoothing out the noise
  (subject of later lecture)

- Simple version: Incorrect hyper-parameters
  Hyper-parameters $=$ parameters that affect algorithm behaviour, including regularisation
- 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%

Hyperparameters

- Parameters = optimised to fit data
- Hyperparameters = set before parameter optimisation
  (pretty arbitrary, but precise meaning for Bayesian models)

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

UNIVERSITY OF BATH

Hyperparameters

- Parameters = optimised to fit data
- Hyperparameters = set before parameter optimisation
  (pretty arbitrary, but precise meaning for Bayesian models)

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

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

Hyperparameters

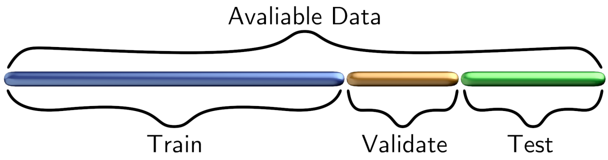
- Parameters = optimised to fit data
- Hyperparameters = set before parameter optimisation
  (pretty arbitrary, but precise meaning for Bayesian models)

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

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

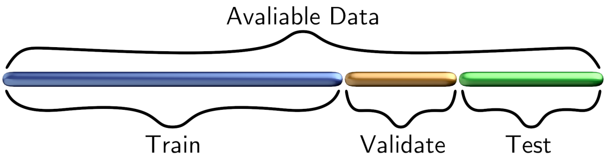
- Introduce a third set: **validation** set
  - **train** – Give to algorithm
  - **validation** – Objective of hyperparameter optimisation
  - **test** – To report final performance

Measuring performance



- How do we decide on split percentages?
  - Train large → Algorithm performs well
  - Validation large → Hyperparameter optimisation performs well, to a limit
  - Test large → Accurate performance estimate, to a limit

Measuring performance

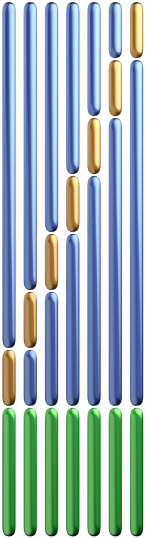


- How do we decide on split percentages?
  - Train large → Algorithm performs well
  - Validation large → Hyperparameter optimisation performs well, to a limit
  - Test large → Accurate performance estimate, to a limit

- Good default: Validation and test small as possible to get reliable estimate, rest on train
- "small as possible" hard to judge however

n-fold

- Validation and test used to make **measurements**
- Can average measurements! (as long as they are independent)

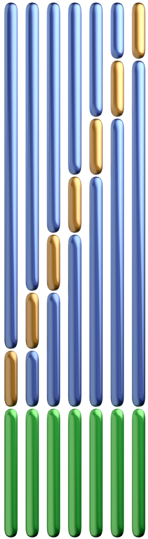


n-fold

- Validation and test used to make **measurements**
- Can average measurements! (as long as they are independent)

- e.g. divide train/validation into 7-fold
  - train: six parts
  - validation: one part

  Train for all seven combinations and report average performance on test



n-fold

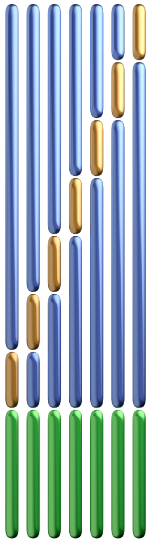
- Validation and test used to make **measurements**
- Can average measurements! (as long as they are independent)

- e.g. divide train/validation into 7-fold
  - train: six parts
  - validation: one part

  Train for all seven combinations and report average performance on test

- n-fold = n $\times$ the computation! Typically $4 \leq n \leq 20$
- General case: All combinations of train/validation/test

- Most extreme: Jackknife resampling
  validation/test sets of size 1; horrifically slow
- In practice mostly not done: time = money



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 for test
- Fast alternative to validation however

Final model

- May train algorithm thousands of times!

  (hyperparameter tuning and n-fold)
- Choice of n is a trade-off between accuracy / time
- Fast computer/cluster/distributed computation really help!
- Final model: Train on entire data set

  (still wise to keep a test set back to sanity check)

Performance?

- What do we actually measure?

  (and hence optimise)

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 wrong
- 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)

Naming more numbers

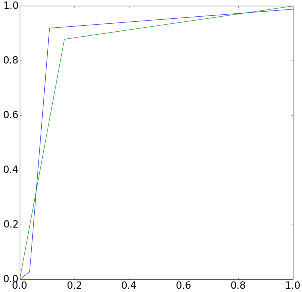
Loads of terms are used (ignore most of them):

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

(many more. . .)

ROC curve

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

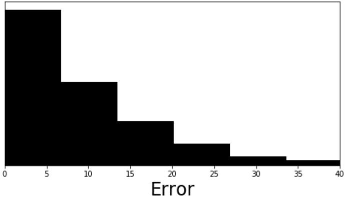


- Threshold sweep. Lets you see the tradeoff – want to be as close to the top left as possible
- True positive rate (x-axis) against false positive rate (y-axis).
- Blue = random forest; Green = linear regression

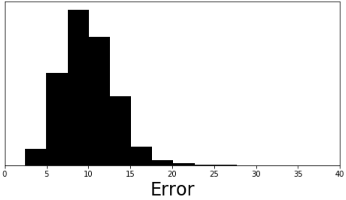
Regression

- Root mean squared error (RMSE) (standard deviation of errors)
- Mean absolute error (MAE)

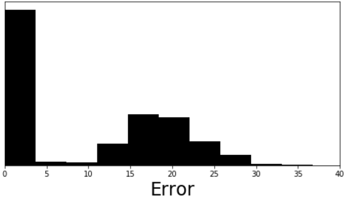
- Max, median, confidence intervals and histograms on absolute errors all have their uses



MAE = 10



MAE = 10



MAE = 10

Performance

- Previous are intermediates
- Need a problem specific function of the confusion matrix (for classification)
- Depending on problem might be better to think in terms of:
  - Cost / Loss
  - Gain
  - Error
  - Risk
- Test complete system!

Group exercise

In small groups discuss how you would measure performance for:

Deciding if a bank should issue a mortgage or not to a customer.	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.

Measurement

If you can't measure it, you can't improve it *(*Peter Drucker*)*

(Obviously false)

UNIVERSITY OF BATH

Measurement

If you can't measure it, you can't improve it *(*Peter Drucker*)*

(Obviously false)

If you can't measure it, you can't apply machine learning to it

(Unfortunately true)

Measurement

If you can't measure it, you can't improve it *(*Peter Drucker*)*

(Obviously false)

If you can't measure it, you can't apply machine learning to it

(Unfortunately true)

- Measurement risk $\rightarrow$ Probability of an incorrect decision due to a measurement
- Measurement is wrong:
  - False positives
  - False negatives

Measurement

If you can't measure it, you can't improve it (*Peter Drucker*)

(Obviously false)

If you can't measure it, you can't apply machine learning to it

(Unfortunately true)

- Measurement risk → Probability of an incorrect decision due to a measurement
- Measurement is wrong:
  - False positives
  - False negatives
- Measurement is right:
  - Measuring the wrong thing
  - Incorrectly interpreting the result

Wrong thing

- How efficient are call center workers?
- Managers measure how quick each call is

University of Bath

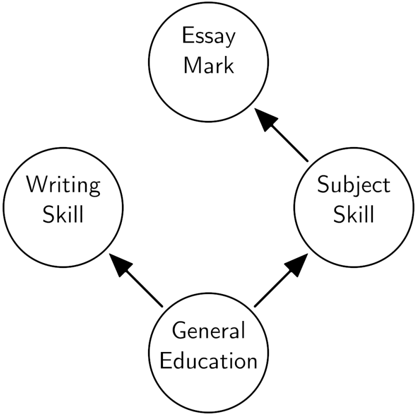
Wrong thing

- How efficient are call center workers?
- Managers measure how quick each call is

- Staff get customers off the phone, rather than solve their problems
- They hang up on hard problems
- They stop showing empathy

Incorrect interpretation

- Build ML system to mark essays
- Features include measures of writing style

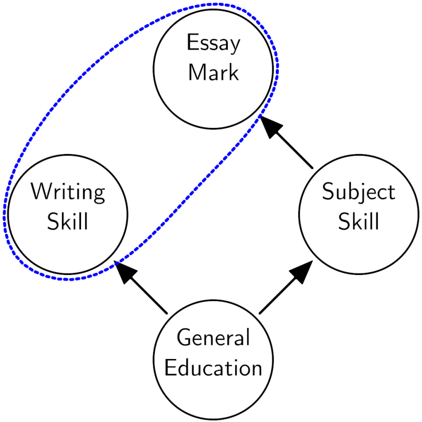


UNIVERSITY OF BATH

Incorrect interpretation

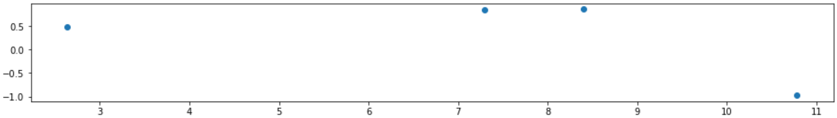
- Build ML system to mark essays
- Features include measures of writing style

- Discover ML primarily uses writing style features
- Example of correlation is not causation

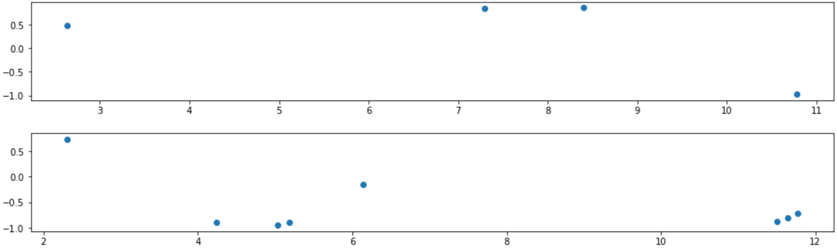


Bad data

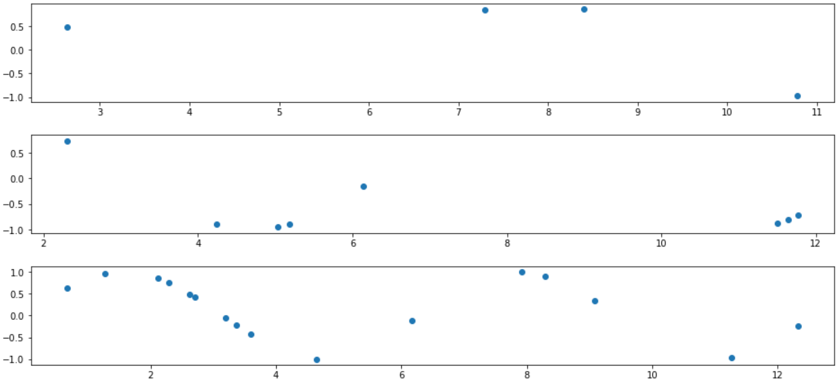
Bad data: Insufficient



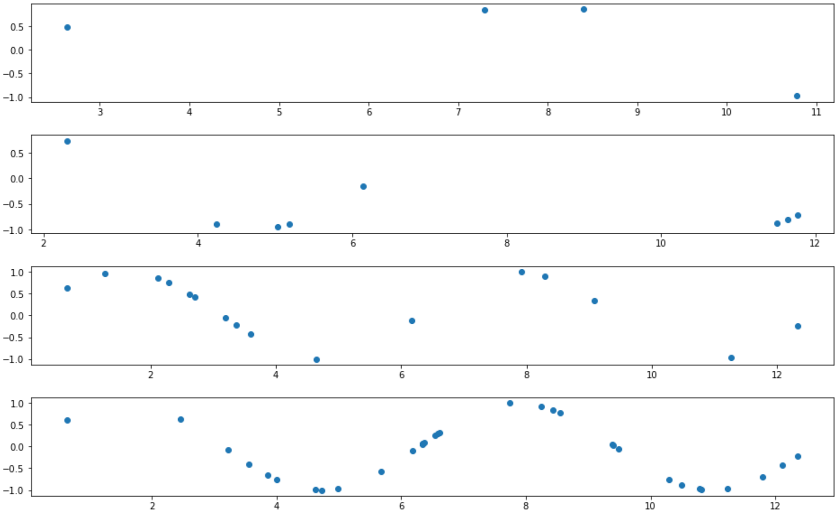
Bad data: Insufficient



Bad data: Insufficient



Bad data: Insufficient



Bad data: Spurious correlation

- 1964: Researcher was spotting M-48 tanks in images
- Got a near perfect score

Bad data: Spurious correlation

- 1964: 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 brightness (b&w so no colour)

- Original paper (probably!): `https://dl.acm.org/citation.cfm?doid=800257.808903`

Bad data: No correlation

- Problem: Estimate when next bus will arrive
- Input:
  - Current height of the fountain
  - Number of purple cars on campus
  - How many bats are in the bat cave
- What's the problem?

Bad data: No correlation

- Problem: Estimate when next bus will arrive
- Input:
  - Current height of the fountain
  - Number of purple cars on campus
  - How many bats are in the bat cave
- What's the problem?
- There is nothing to learn – no correlation – it's impossible!

(should output average wait time)

UNIVERSITY OF BATH

Bad data: Unbalanced

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

  ...and often does

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

Bad data: Selection bias

- WW2: US military wanted add armour to bombers
- Initial idea: Add it where there are holes on the returning bombers

Bad data: Selection bias

- WW2: US military wanted add armour to bombers
- Initial idea: Add it where there are holes on the returning bombers

- Abraham Wald (statistician) 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!

Bad data: Selection bias

- WW2: US military wanted add armour to bombers
- Initial idea: Add it where there are holes on the returning bombers
- Abraham Wald (statistician) 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: Runtime mismatch

(hypothetical scenario)

- Train system to identify car model
- Performs well on test set

Bad data: Runtime mismatch

(hypothetical scenario)

- Train system to identify car model
- Performs well on test set

- Trained and tested in UK...

  ...deployed in Cuba
- Different cars, so it fails

Bad data: Missing context

- Hospital must decide to admit or send home patients with pneumonia
- Uses ML trained on survival rate
- Recommends sending patients with asthma home. . .

. . . which will probably kill them

Bad data: Missing context

- Hospital must decide to admit or send home patients with pneumonia
- Uses ML trained on survival rate
- Recommends sending patients with asthma home...

  ...which will probably kill them

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

- Reported in: `http://people.dbmi.columbia.edu/noemie/papers/15kdd.pdf`

Bad data: Biased data I

- Florida: Dentencing influenced by predicted reoffending rate
- ML to predict, trained on historic judgements

Bad data: Biased data I

- Florida: Dentencing influenced by predicted reoffending rate
- ML to predict, trained on historic judgements

- Turns out the judges are racist
- ML replicates their racism

- This: 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

Bad data: Biased data II

(2018-10-10)

- Amazon has too many job applications
- ML system sorts best to worst to save time

Bad data: Biased data II

(2018-10-10)

- Amazon has too many job applications
- ML system sorts best to worst to save time

- Learned to dislike women
- Trained on past applicants. . . which are mostly male
- Spurious correlation as well as bad data
- Article: `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G`

Detecting bad data

- Visualise (but be careful – a bad visualisation can be misleading)
- Use multiple performance metrics

Detecting bad data

- Visualise (but be careful – a bad visualisation can be misleading)
- Use multiple performance metrics

- Test for failure scenarios:
  e.g. Verify that changing the gender of a CV doesn't change the rating

Summary

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

Further reading

- Blog breaking down why a medical data set is useless: https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/
- "Concrete Problems in AI Safety"
  by **Amodei, Olah, Steinhardt, Christiano, Schulman and Mane**
  https://arxiv.org/pdf/1606.06565.pdf