ERHVERVSAKADEMI
AARHUS

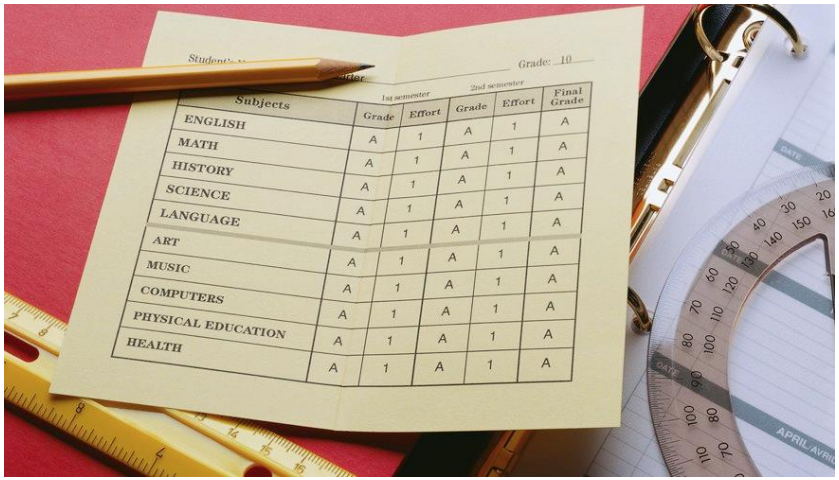
Machine Learning – Week 7

Sila, Oct 7th 2025.

Welcome - Todays Program.

- The Titanic Competition
- Model Optimization. Model evaluation. Spotting anomalies.
- Model Optimization. Underfit and overfit.
- Model Optimization. Confusion Matrix (revisited).
- Model Optimization. Feature engineering (revisited).
- Model Optimization. Datacompression.
  Principal Component Analysis(PCA) for dimensionality reduction.
- A machine learning pipeline. Putting it all together.

Handins. Please Remember. Will say this many times…



- Do the work.
- Individual reports.
- Plan ahead.

Handins. Please Remember:

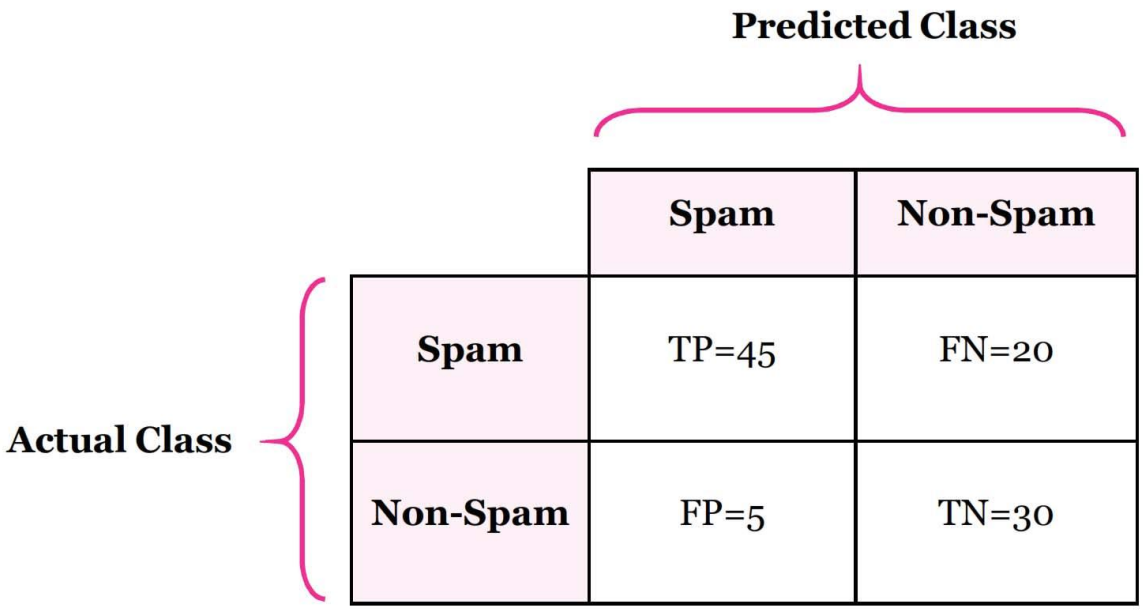
Notes for later (for ITA).

- Handin a zip file that includes **pdf report, and .py files.**
(typical size 1 - 10 MBs ,,, not 300 MBs, and
definitely not GBs..)

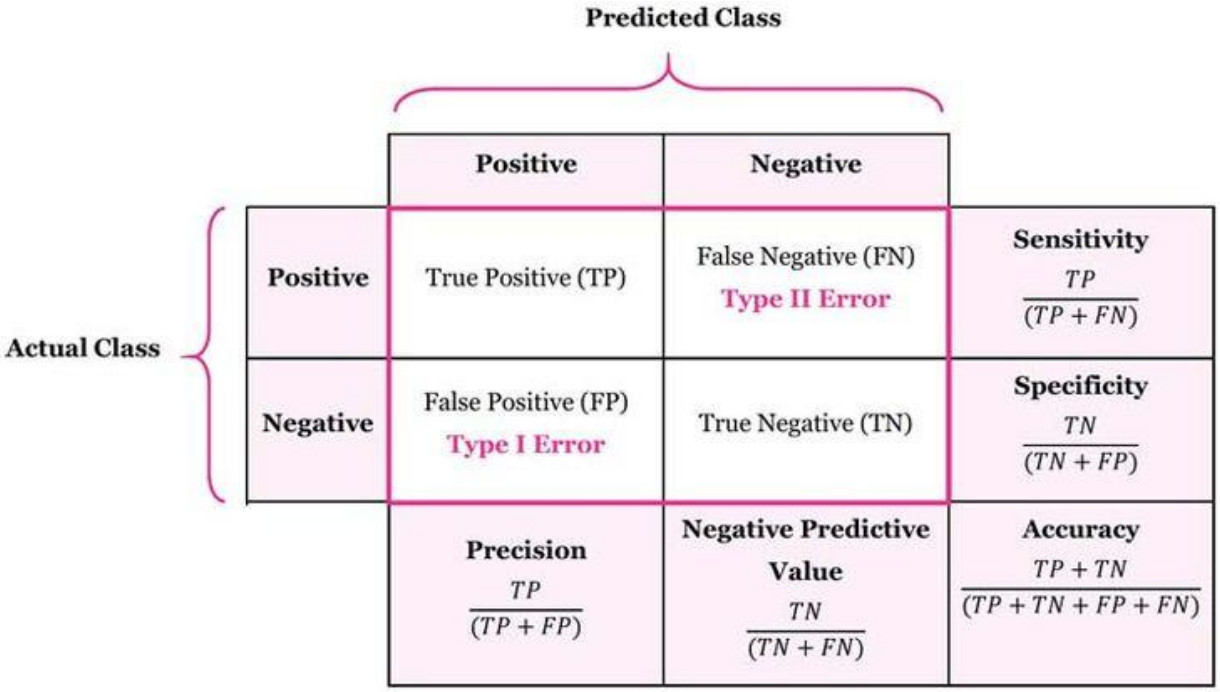
- Individual handin – You can talk with classmates
in the process, also the code, but you handin individual code
and reflections. So, remember, *you do the work*! *Your handin* !

- Size of report – not super important… Just give
results and reflections.

Machine Learning – Confusion Matrix. Recap.



Machine Learning – Confusion Matrix. Recap.

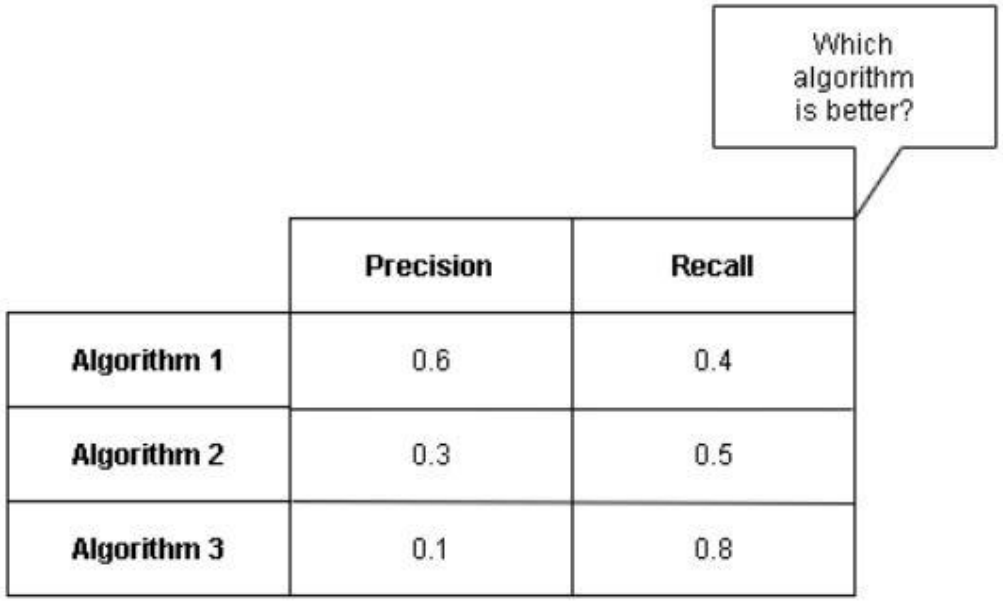


In groups of 2-3.

Recap.

What is accuracy?
What is precision?
What is recall?

Machine Learning – Confusion Matrix. Recap.



	Precision	Recall
Algorithm 1	0.6	0.4
Algorithm 2	0.3	0.5
Algorithm 3	0.1	0.8

Discuss:Which algorithm is best?

Machine Learning – Confusion Matrix. Recap.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precison} + \text{Recall}}$$

F1 Score Formula (Image Source: Author)

Having a precision or recall value as 0 is not desirable and hence it will give us the F1 score of 0 (lowest). On the other hand, if both the precision and recall value is 1, it'll give us the F1 score of 1 indicating perfect precision-recall values. All the other intermediate values of the F1 score ranges between 0 and 1.

Which algorithm is best?

Titanic Competition

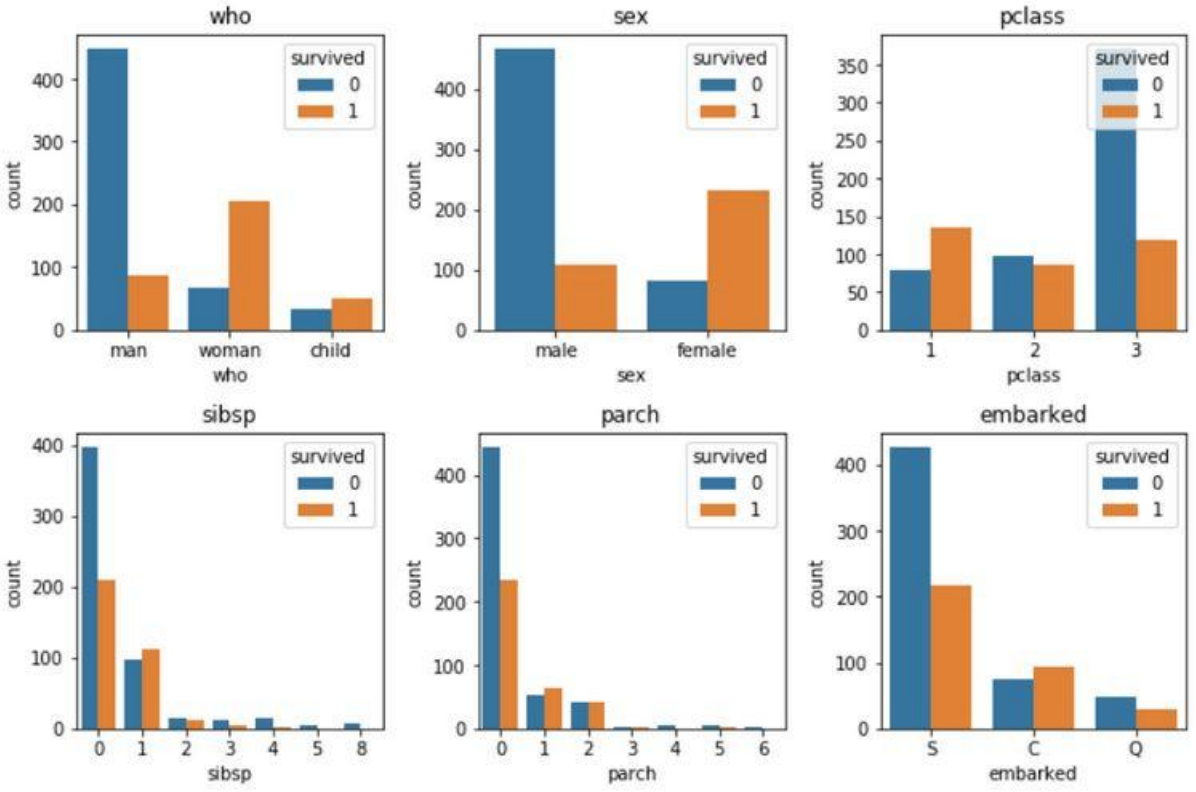
- The Titanic Competition

  In groups of 2: How and what did you do?

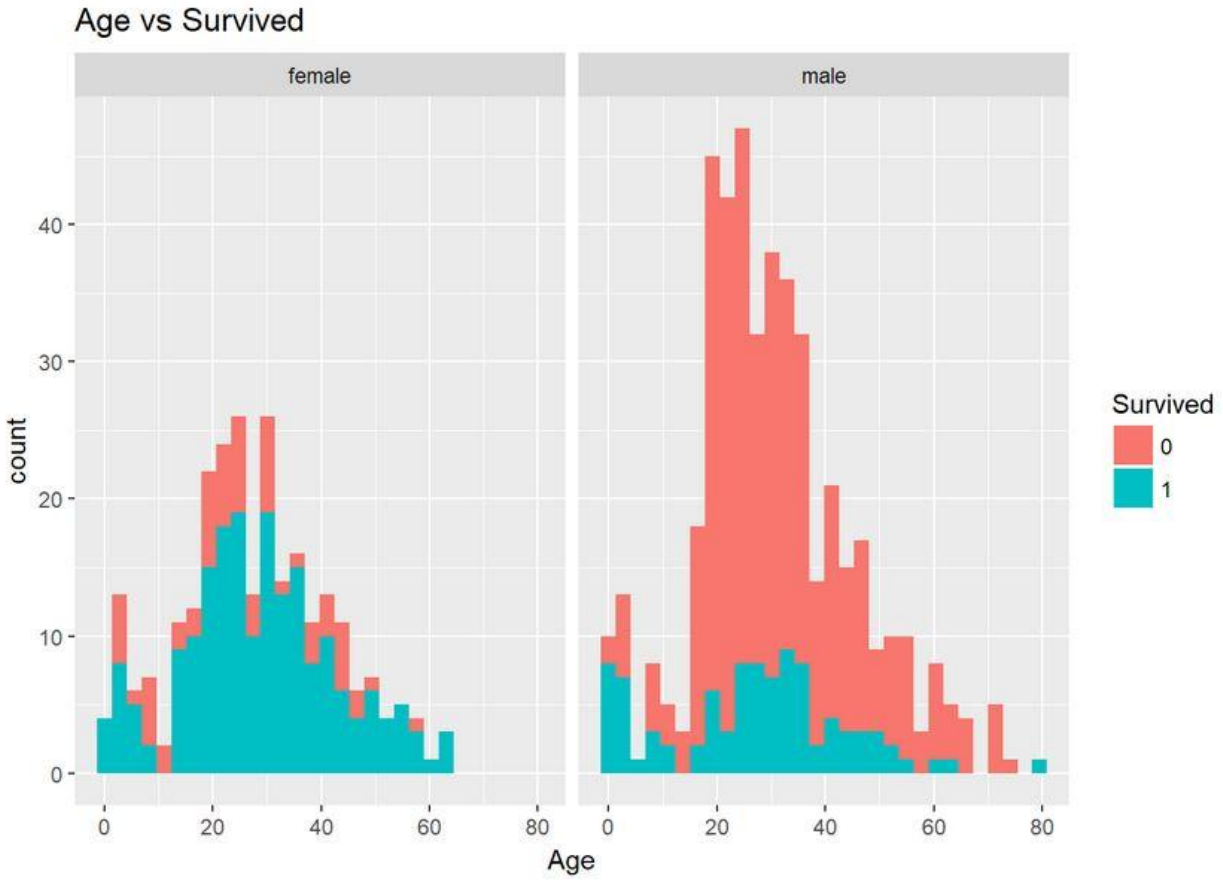
  Show solutions to each other, and discuss, in detail, what you have done, and what you would have liked to do, given more time.
  What was the result?

  Afterwards:
  We discuss it in class.

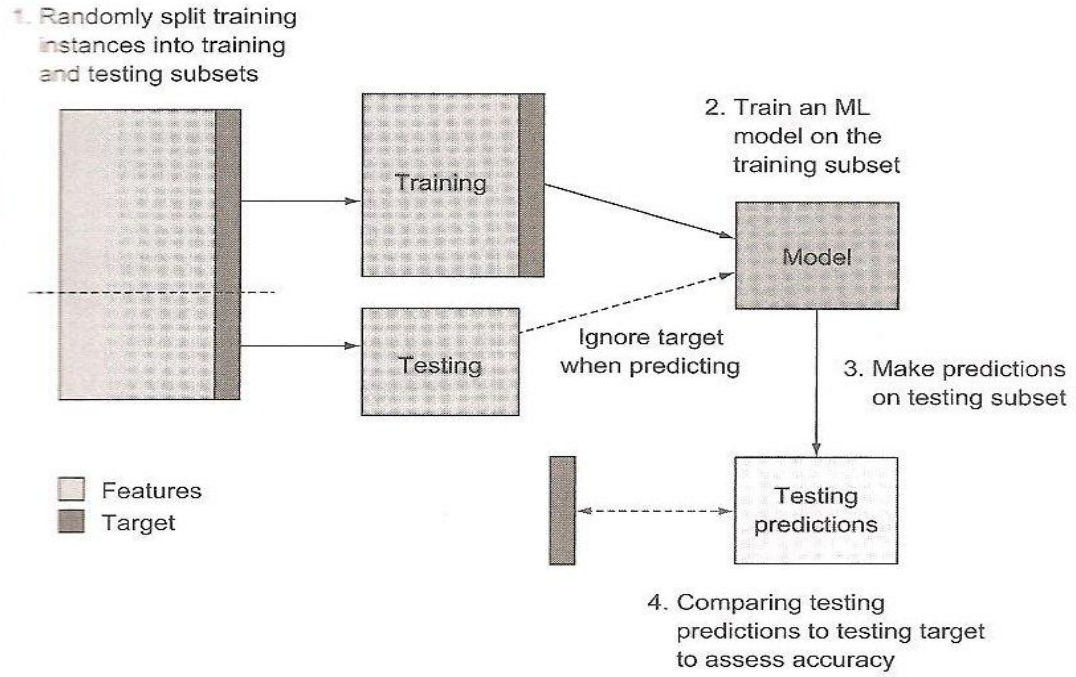
Titanic Competition



Titanic Competition



Machine Learning – Model optimization, model evaluation.



Model evaluation. Cross validation.

Use part of the training data to evaluate how good the model will be on new data.

See p. 83 in "Real World ML" by Brink et al.

Machine Learning – Model optimization. A closer look at the Titanic dataset, cross validation:

Target column

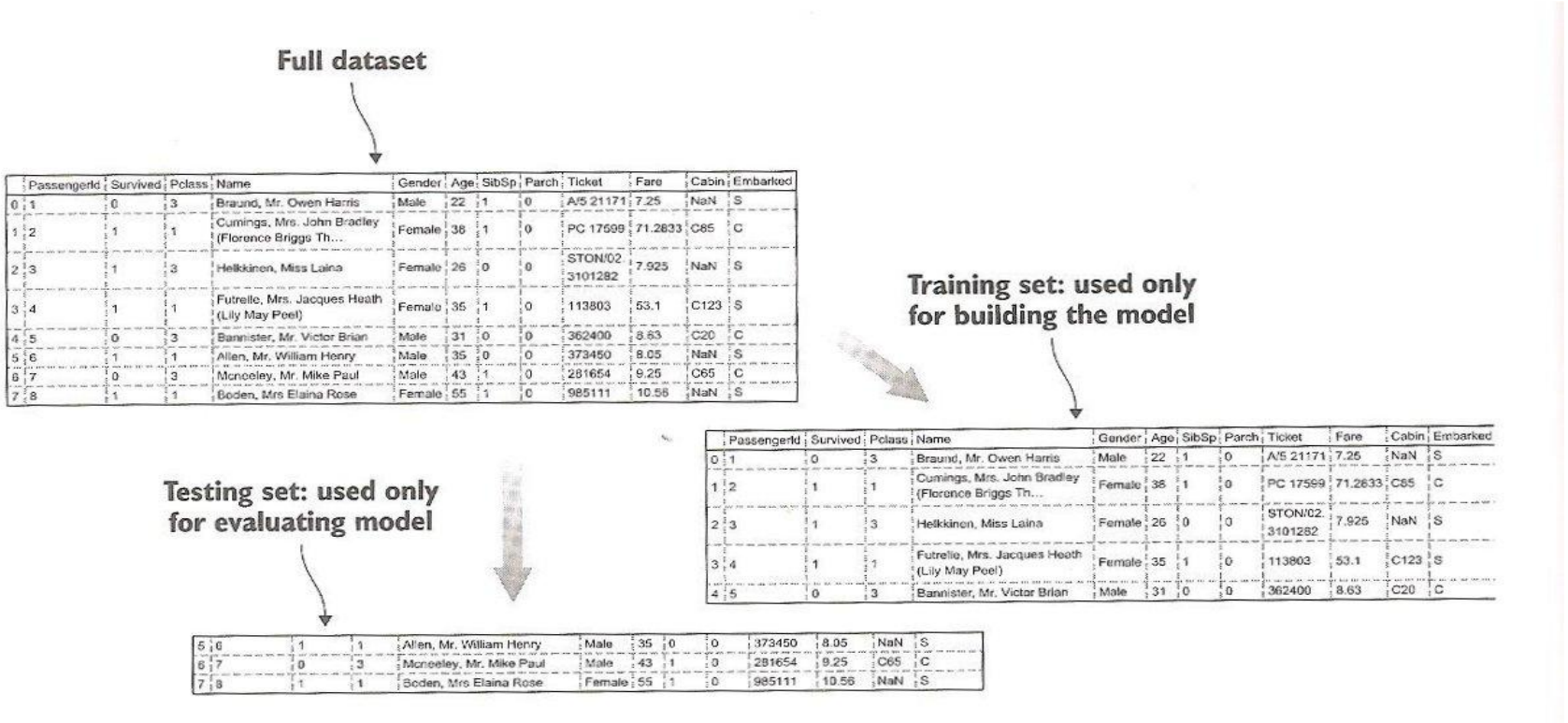
The target column indicates whether a passenger survived the sinking of the ship or died.

	PassengerId	Survived	Pclass	Name	Gender	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	Male	22	1	0	A/5 21171	7.25	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th…	Female	38	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Helkkinen, Miss Laina	Female	26	0	0	STON/02. 3101282	7.925	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	Female	35	1	0	113803	53.1	C123	S
4	5	0	3	Allen, Mr. William Henry	Male	35	0	0	373450	8.05	NaN	S

The first five rows of the Titanic Passengers dataset

For more comments about the dataset -
See p. 87 in "Real World ML" by Brink et al.

Machine Learning – Model optimization. A closer look at the Titanic dataset, cross validation:



	PassengerId	Survived	Pclass	Name	Gender	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	Male	22	1	0	A/5 21171	7.25	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	Female	38	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Helkkinen, Miss Laina	Female	26	0	0	STON/02 3101282	7.925	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	Female	35	1	0	113803	53.1	C123	S
4	5	0	3	Bannister, Mr. Victor Brian	Male	31	0	0	362400	8.63	C20	C
5	6	1	1	Allen, Mr. William Henry	Male	35	0	0	373450	8.05	NaN	S
6	7	0	3	Mcneeley, Mr. Mike Paul	Male	43	1	0	281654	9.25	C65	C
7	8	1	1	Boden, Mrs Elaina Rose	Female	55	1	0	985111	10.56	NaN	S

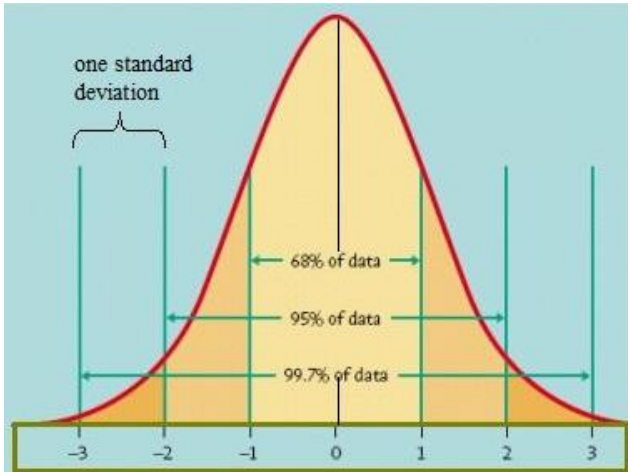
For comments about the dataset -
See p. 87 in "Real World ML" by Brink et al.

Exercise 1. Spotting anomaly.

In groups of two, how can you spot anomaly in a dataset (or: might be able to spot) ?

What can you do? What should you do?

Machine Learning – Model optimization. Looking at the data. Anomaly detection…



In the case of normally distributed data,
the three sigma rule means that roughly 1 in 22 observations will differ by twice the standard deviation or more from the mean, and 1 in 370 will deviate by three times the standard deviation.

Machine Learning – Model optimization. Anomaly detection.

Anomaly detection algorithm

1. Choose features x_i that you think might be indicative of anomalous examples.
2. Fit parameters $\mu_1, \ldots, \mu_n, \sigma_1^2, \ldots, \sigma_n^2$

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)}$$

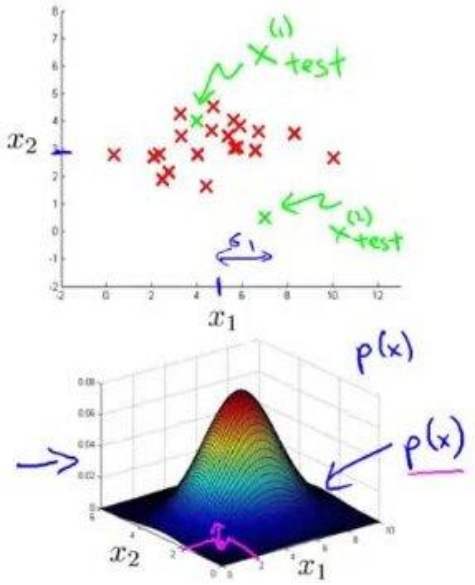
$$\sigma_j^2 = \frac{1}{m}\sum_{i=1}^{m}(x_j^{(i)} - \mu_j)^2$$

3. Given new example x, compute $p(x)$:

$$p(x) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}\right)$$

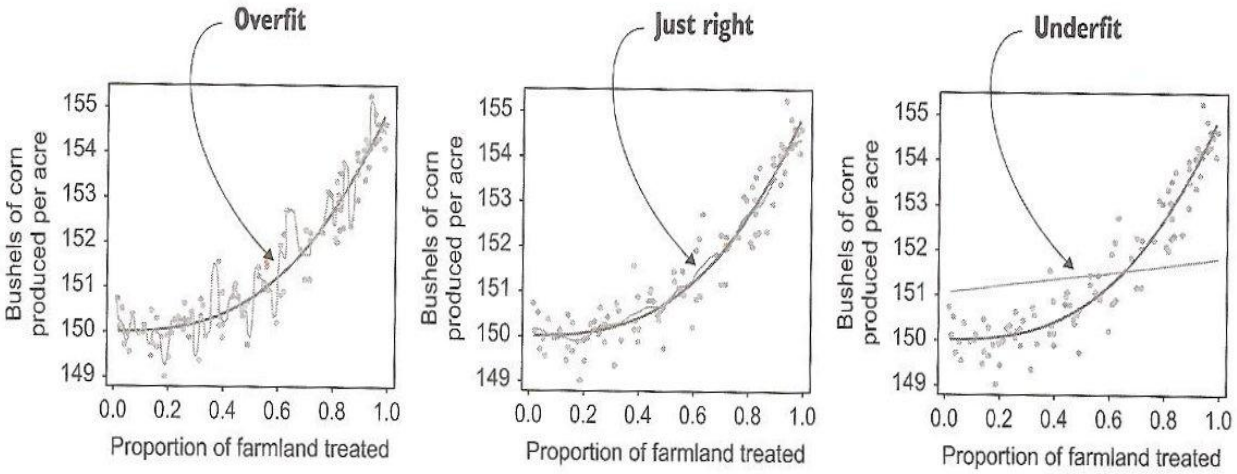
Anomaly if $p(x) < \varepsilon$

Anomaly detection example



Andrew Ng, Coursera.

Machine Learning – Model optimization



Finding the right model.
Three fits of a model to a ”corn-production” training set.

See p. 80 in "Real World ML" by Brink et al.

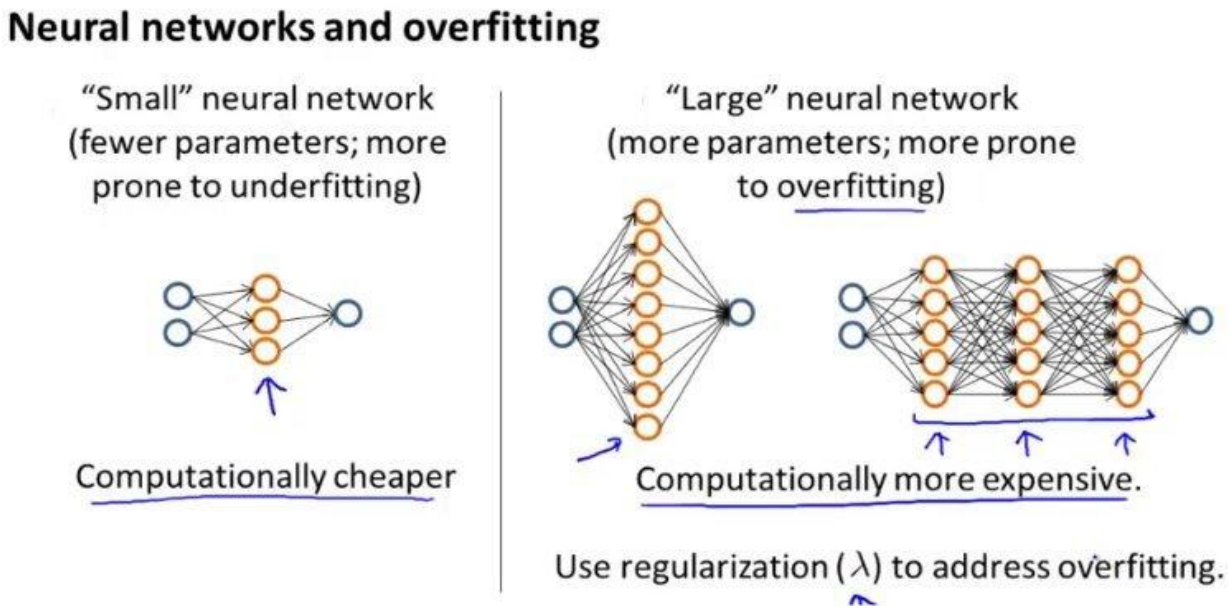
Exercise 2. Underfit and overfit.

In groups of 2-3.

if your model is underfitting or overfitting

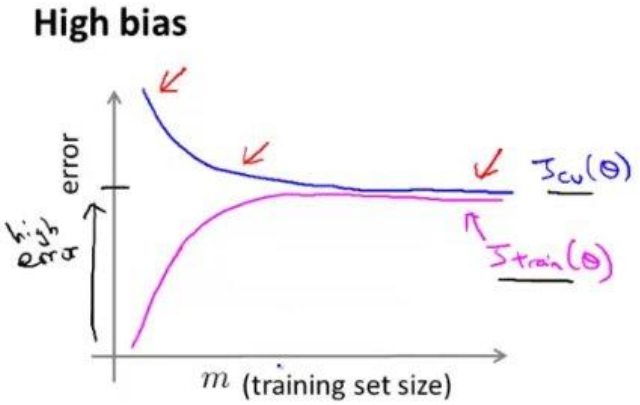
- What might be able to make the model better? What could you try out?

Machine Learning – Model optimization

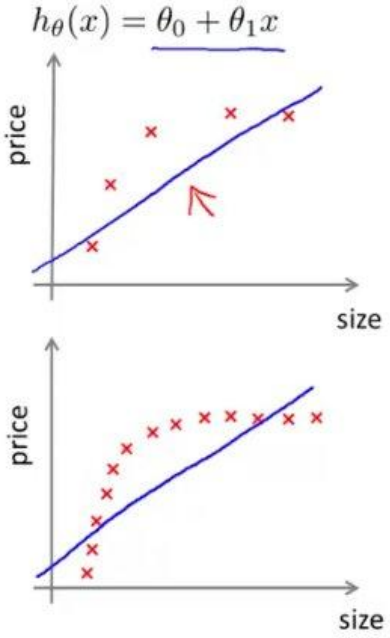


Andrew Ng. Underfitting and overfitting. From Coursera.

Machine Learning – Model optimization



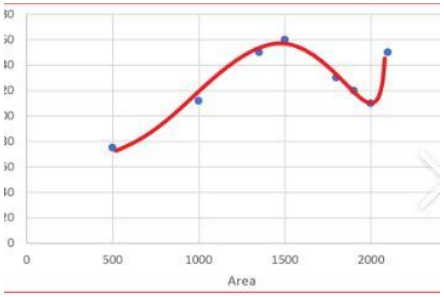
If a learning algorithm is suffering from high bias, getting more training data will not (by itself) help much.



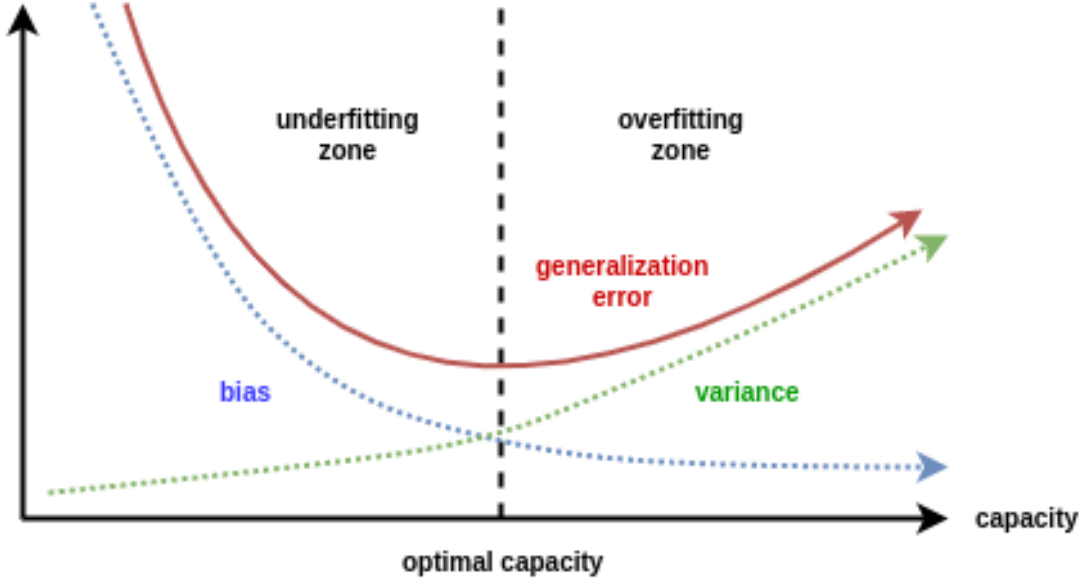
Andrew Ng. Underfitting and overfitting. From Coursera.

Machine Learning – Model optimization

Variance refers to the **sensitivity of the learning algorithm to the specifics of the training data**, e.g. the noise and specific observations.



High Variance – overfit

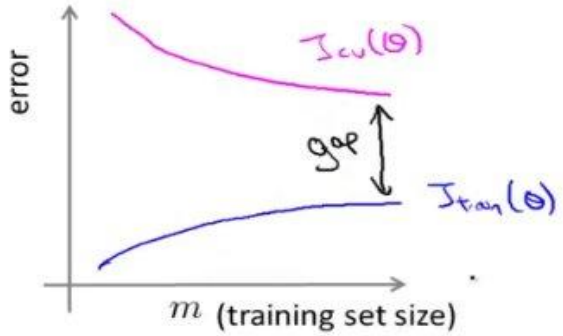


Andrew Ng. Underfitting and overfitting. From Coursera.

Variance is the variability in the model prediction—how much the ML function can adjust depending on the given data set.

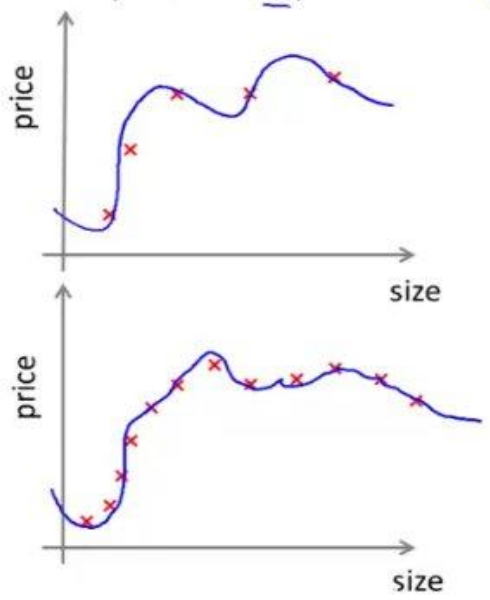
Machine Learning – Model optimization

High variance



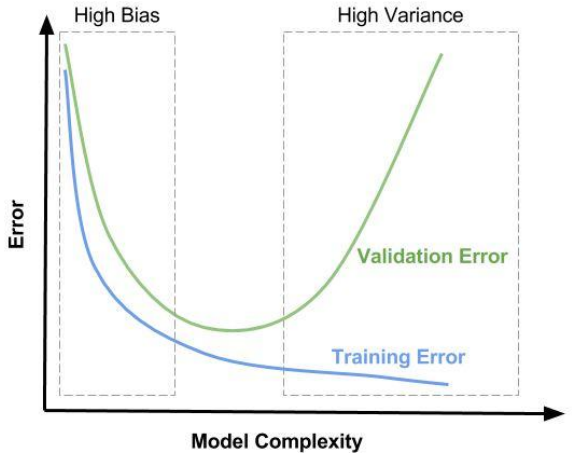
If a learning algorithm is suffering from high variance, getting more training data is likely to help.

$h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_{100} x^{100}$
(and small λ)



Andrew Ng. Underfitting and overfitting. From Coursera.

Machine Learning – Model optimization



The region on the left, where both training and validation errors are high, is the region of high bias. On the other hand, the region on the right where validation error is high, but training error is low is the region of high variance. We want to be in the sweet spot in the middle.

Variance quotes:

Error due to variance is the amount by which the prediction, over one training set, differs from the expected value over all the training sets. In machine learning, different training data sets will result in a different estimation. But ideally it should not vary too much between training sets. However, if a method has high variance then small changes in the training data can result in large changes in results.

The variance is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).

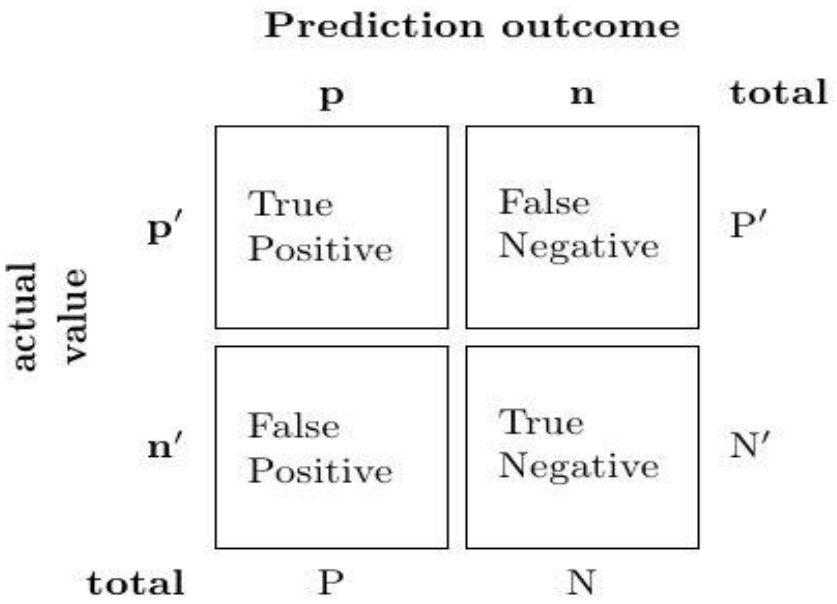
Machine Learning – Model optimization

Our decision process can be broken down as follows:

- **Getting more training examples:** Fixes high variance
- **Trying smaller sets of features:** Fixes high variance
- **Adding features:** Fixes high bias
- **Adding polynomial features:** Fixes high bias
- **Decreasing λ:** Fixes high bias
- **Increasing λ:** Fixes high variance.

Andrew Ng. Underfitting and overfitting. From Coursera.

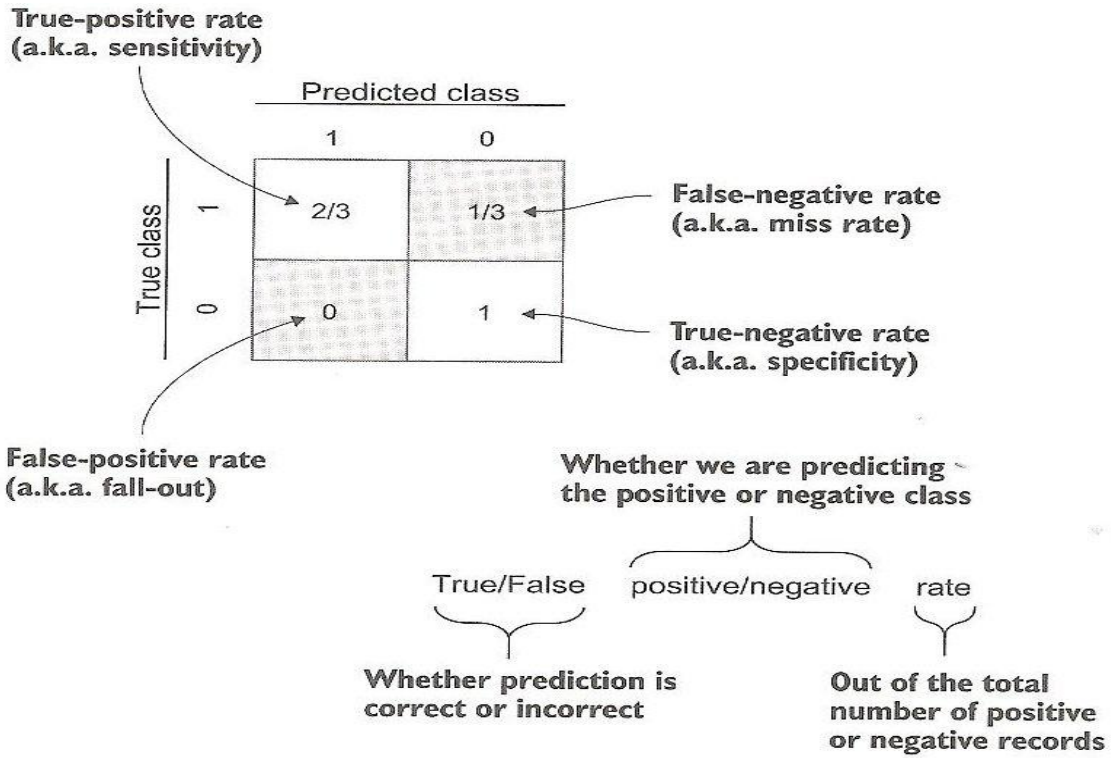
Machine Learning – Confusion Matrix. Selecting a good model. ROC Curves.



For more about the confusion matrix.
See p. 90 in "Real World ML" by Brink et al.

Machine Learning – Confusion Matrix.

Selecting a good model. ROC Curves.



For more about the confusion matrix.
See p. 90 in "Real World ML" by Brink et al.

Machine Learning – Confusion Matrix.
Selecting a good model. ROC Curves.

	Not Pregnant	Pregnant
Positive Test Result	False Positive	True Positive
Negative Test Result	True Negative	False Negative

		True diagnosis		
		Positive	Negative	Total
Screening test	Positive	a	b	$a+b$
	Negative	c	d	$c+d$
	Total	$a+c$	$b+d$	N

Predicted Class	True Outcome : Patients have Disease A	
	Positive (Patients have disease A)	Negative (Patients do not have disease A)
Positive (Patients have disease A)	True Positives	False Positives (Patients wrongly identified to have disease A)
Negative (Patients do not have disease A)	False Negatives (Patients have been left out from treatment for Disease)	True Negatives

For more about the confusion matrix.
See p. 90 in "Real World ML" by Brink et al.

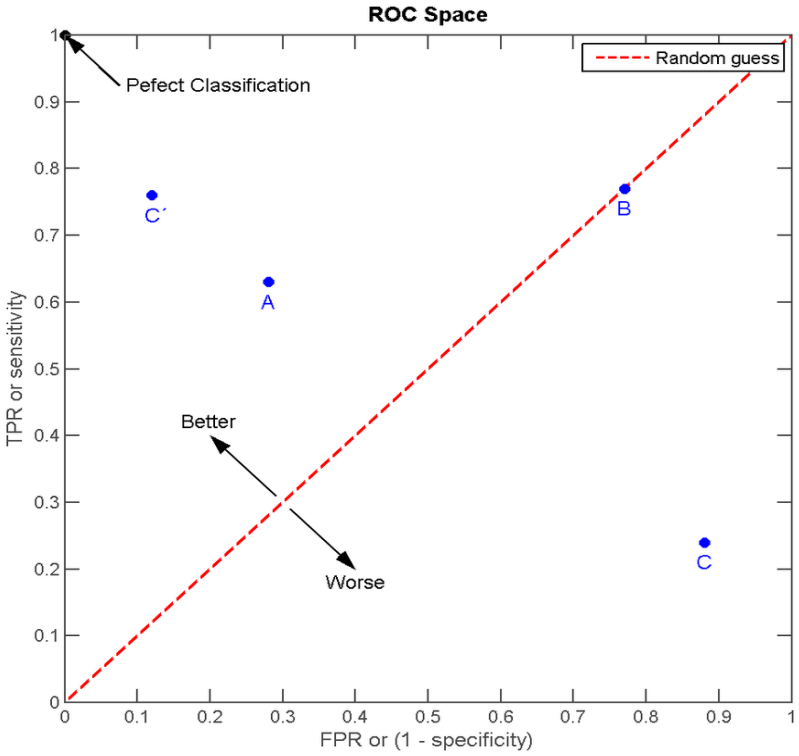
Machine Learning – Confusion Matrix.

Selecting a good model. ROC Curves.

Predicted Class	True Outcome : Customers Default or Not	
	Positive (or Good)	Negative (Bad)
Positive (or Good)	True Positives	False Positives (Type I Error)
Negative (Bad)	False Negatives (Type II Error)	True Negatives

For more about the confusion matrix.
See p. 90 in "Real World ML" by Brink et al.

Model optimization. ROC curves (receiver operating characteristics). ROC Curves.



A			B			C			C'		
TP=63	FP=28	91	TP=77	FP=77	154	TP=24	FP=88	112	TP=76	FP=12	88
FN=37	TN=72	109	FN=23	TN=23	46	FN=76	TN=12	88	FN=24	TN=88	112
100	100	200	100	100	200	100	100	200	100	100	200

For more about the confusion matrix.
See p. 91 in "Real World ML" by Brink et al.

Model optimization. A closer look at the Titanic dataset, thresholds. ROC Curves.

Output from classifier: class probabilities

	Survived	Died
15	0.092	0.908
16	0.904	0.096
17	0.646	0.354
18	0.740	0.260
19	0.460	0.540

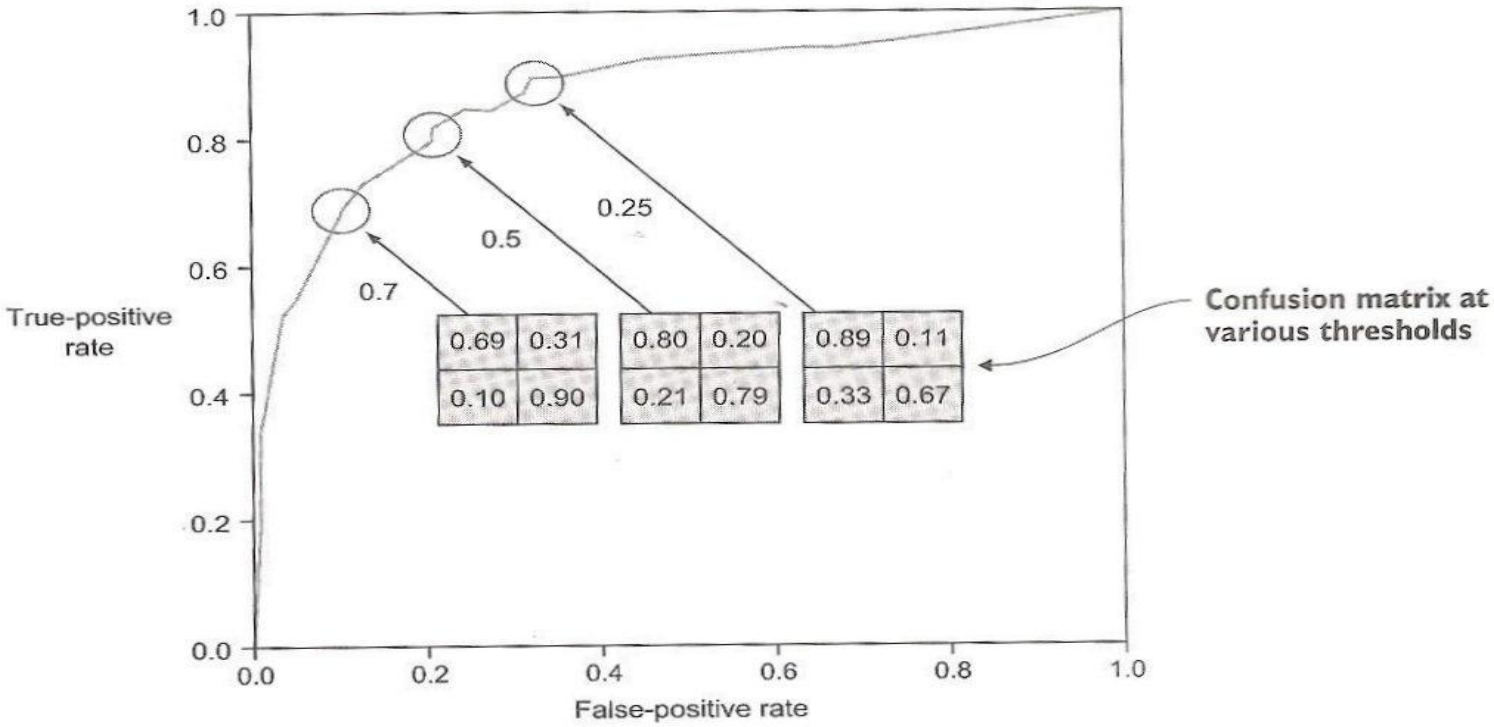
Sorted probabilities

	Survived	Died
308	0.705	0.295
215	0.703	0.297
217	0.700	0.300
54	0.698	0.302
169	0.698	0.302

Threshold: "survived" probabilities > 0.7

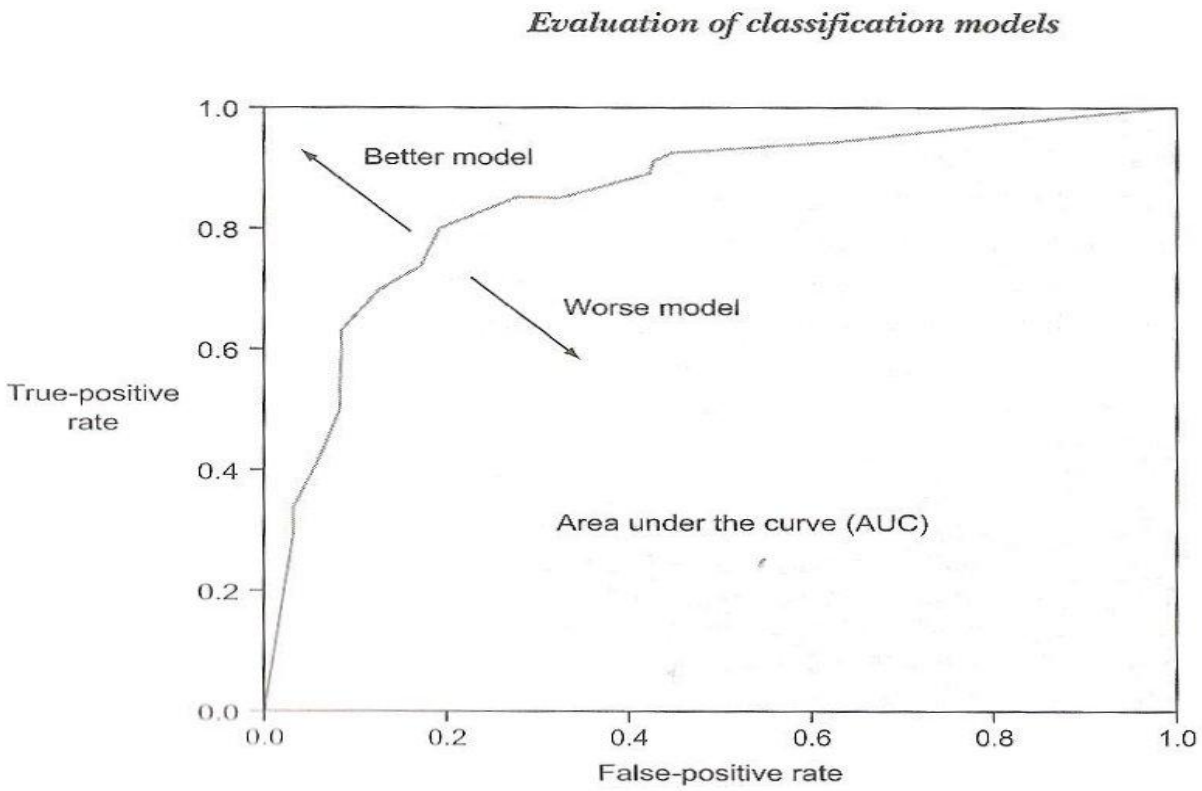
For more about the confusion matrix.
See p. 91 in "Real World ML" by Brink et al.

Model optimization. Roc curves. thresholds.



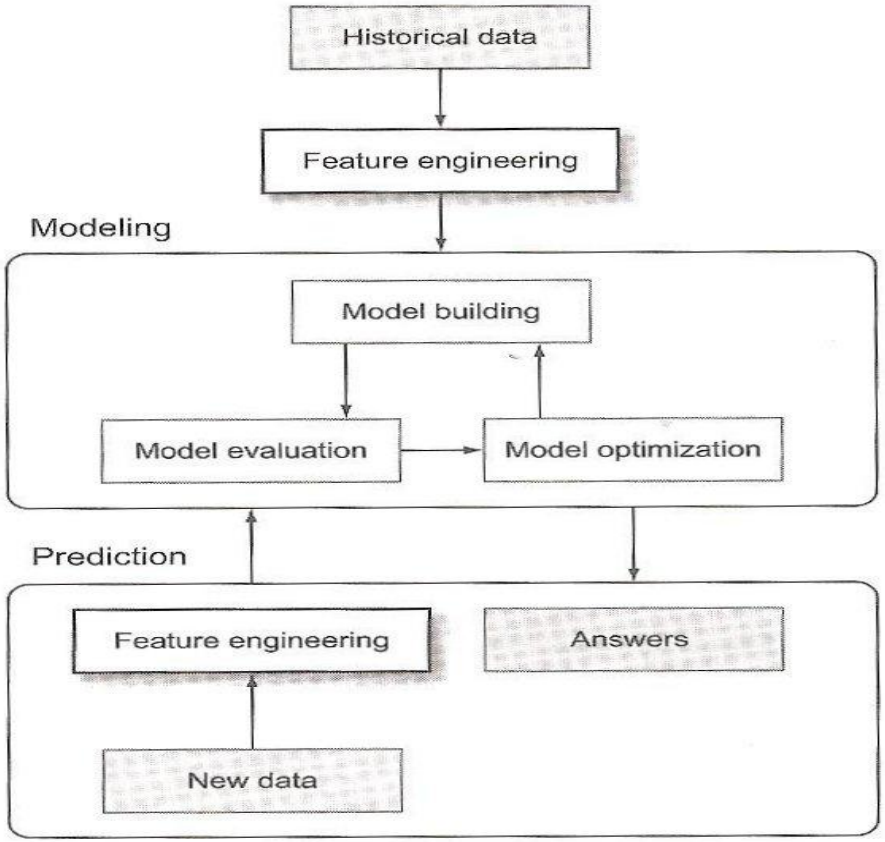
For more about the confusion matrix.
See p. 91 in "Real World ML" by Brink et al.

Model optimization. Roc curves. Thresholds.



For more about the confusion matrix.
See p. 91 in "Real World ML" by Brink et al.

Model optimization. Model building. Model evaluation. Feature engineering.



Model optimization. Model building.
Feature evaluation/Engineering. Model evaluation.

datetime_hour_of_day	datetime_day_of_week	datetime_day_of_month	datetime_day_of_year	datetime_month_of_year
13	4	26	300	10
13	4	26	300	10
13	4	26	300	10
13	4	26	300	10
13	4	26	300	10

datetime_minute_of_hour	datetime_second_of_minute	datetime_year	datetime_quarter_of_year	datetime_week_of_year
30	0	2012	4	43
30	0	2012	4	43
30	0	2012	4	43
30	0	2012	4	43
30	0	2012	4	43

Figure 5.4 Additional date-time columns extracted from the timestamp column for the event-recommendation dataset

Feature engineering.
See p. 112 in "Real World ML" by Brink et al.

Model optimization. Model building.
Model is too complex?? Datacompression. PCA.

Unsupervised transformations.

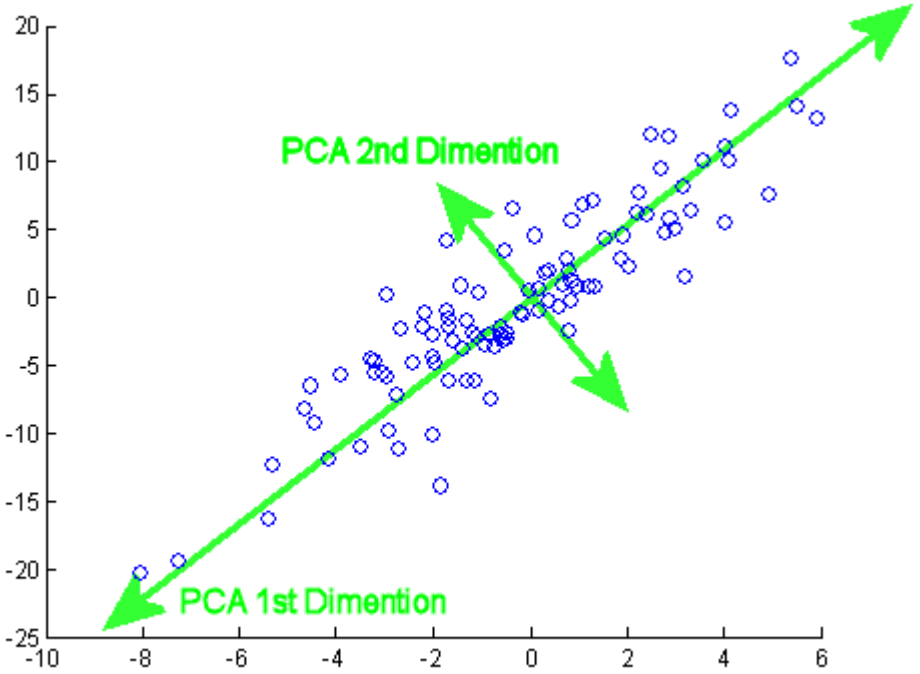
A high-dimensionality representation of data, consisting of many features,
Is transformed to a new representation that summarizes the essential characteristics with fewer features...

A common application is reduction to two dimensions
for better visualization.

PCA (Principal Component Analysis)
– Main idea.

Model optimization. Model building.

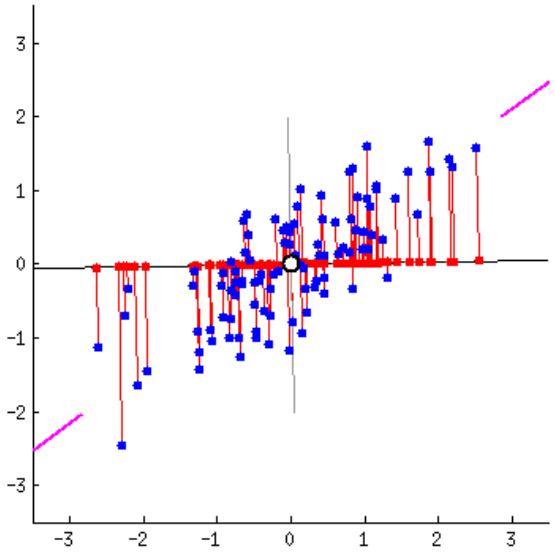
Model is too complex?? Datacompression. PCA.



PCA – Main idea.

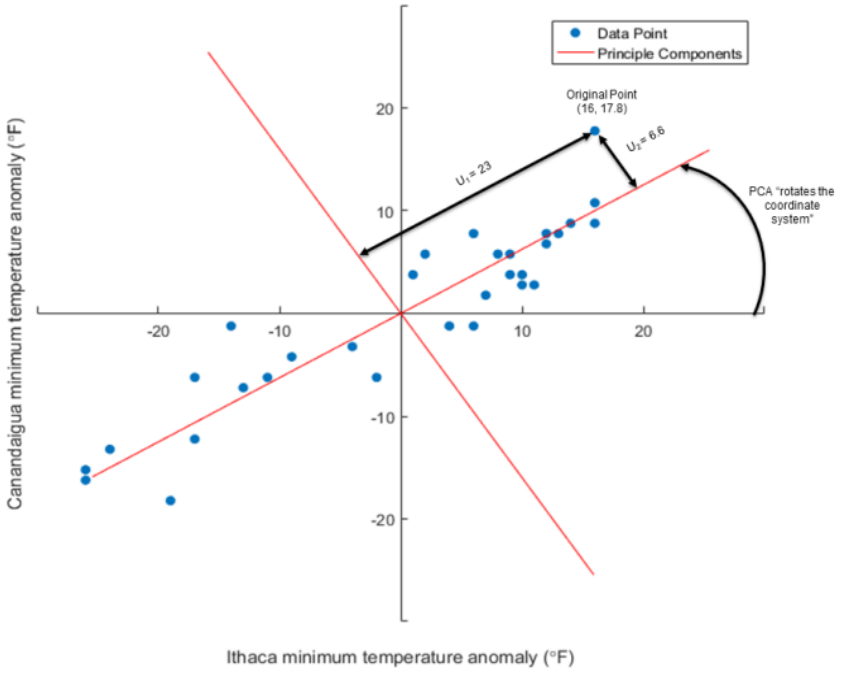
Model optimization. Model building.

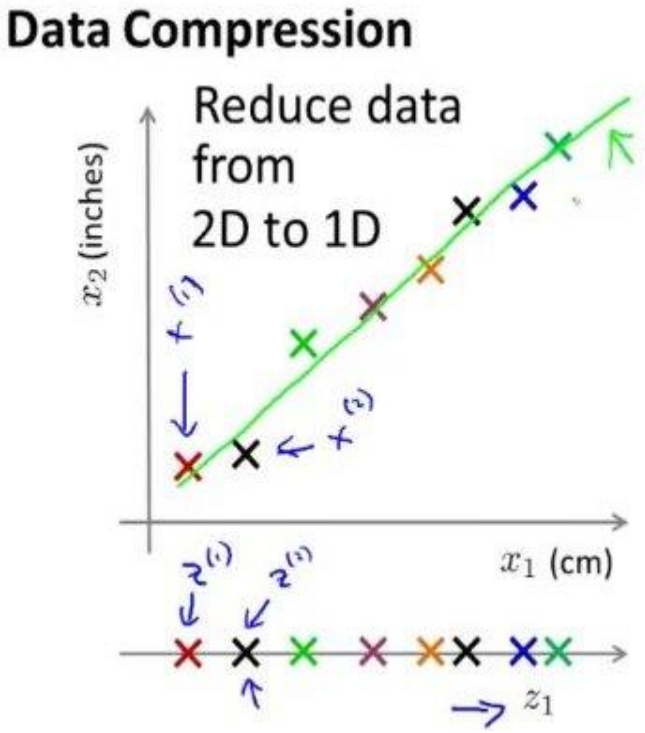
Model is too complex?? Datacompression. PCA.



PCA – Main idea.

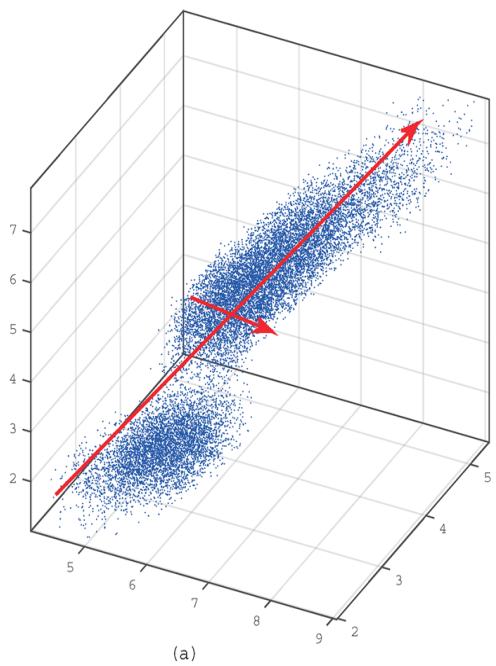
Model optimization. Model building. Datacompression. PCA.

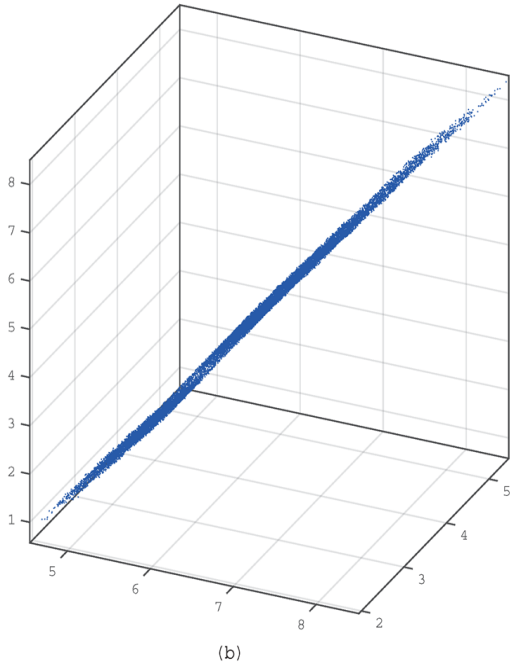


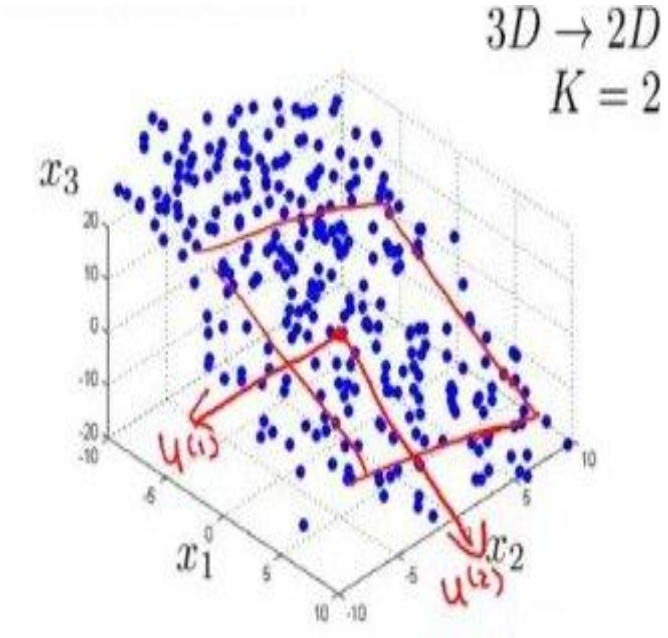


PCA Example. PCA - ”rotates the coordinate system”

Model optimization. Model building. Datacompression. PCA.







PCA in a plane.

Model optimization. Model building. Datacompression. PCA.

Exercise 3:

Why is data compression useful in machine learning?

Model optimization. Model building. Datacompression. PCA.

Application of PCA

- Reduce memory/disk needed to store data
- Speed up learning algorithm

- Visualization:
k = 2 or k = 3

Model optimization. Model building. Datacompression. PCA.

Application of PCA

- To avoid overfitting?

Not good…!

The reason is that PCA is not aware of the labels in your training set,
PCA just throws away some information
but it doesn't know what is useful and
what is not for the purpose of classification.

Overfitting happens when the model learns on noise as well as the impt features. So pca doesn't really do anything in that regard, in the sense that it doesn't really say which features are most important for the model, just which features showed the biggest variance.

Model optimization. Model building. Datacompression. PCA.

Bad use of PCA: To prevent overfitting

→ Use $z^{(i)}$ instead of $x^{(i)}$ to reduce the number of features to $k < n$. — 1000 — 10000

Thus, fewer features, less likely to overfit.

Bad!

This might work OK, but isn't a good way to address overfitting. Use regularization instead.

$$\min_\theta \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

Andrew Ng, Coursera.

Model optimization. Model building. Datacompression. PCA.

fewer features, less likely to overfit.

Bad!

Sometimes: Using PCA also seem to reduce the chance of overfitting your model by eliminating features with high correlation.

But remember:

Trying to decide if a component of a PCA shall be retained, or not:

Answer: Only get rid of data that is not valuable...

I recently had optical spectroscopy data, where > 99% of the total variance of the raw data was due to changes in the background light (spotlight more or less intense on the measured point, fluorescent lamps switched on/off, more or less clouds before the sun). After background correction with the optical spectra of known influencing factors (extracted by PCA on the raw data; extra measurements taken in order to cover those variations), the effect we were interested in showed up in PCs 4 and 5. PCs 1 and 3 where due to other effects in the measured sample, and PC 2 correlates with the instrument tip heating up during the measurements.

Model optimization. Model building. Datacompression. PCA.

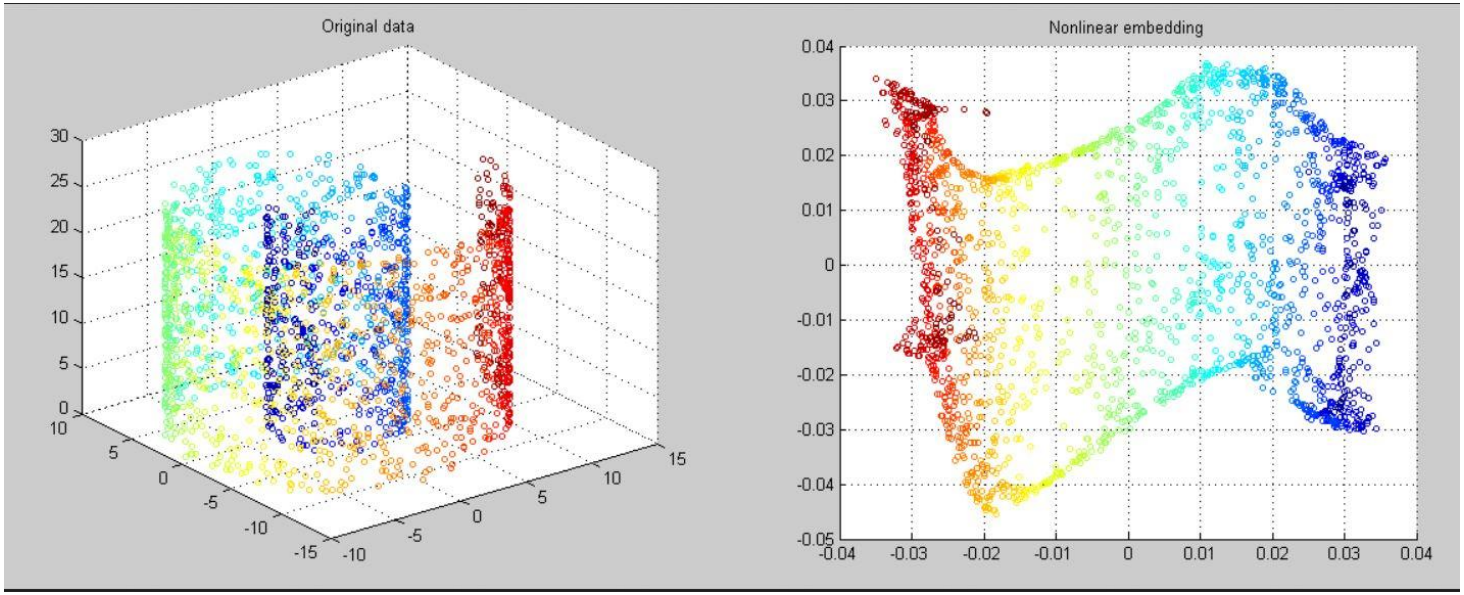
PCA aims to find the directions of maximum variance in high-dimensional data and projects it onto a new subspace with equal or fewer dimensions than the original one.

PCA is unsupervised and *projects the data on the direction with the most variance regardless of the label of the data*. This means that it will *most of the time lose some critical information* that is (might be) fundamental for classification.

Note that people sometimes use PCA to prevent overfitting since fewer features implies that the model is less likely to overfit. While PCA may work in this context, it is not a good approach! The reason is that PCA is not aware of the labels in your training set, PCA just throws away some information but it doesn't know what is useful and what is not (for the purpose of classification).

Principal component analysis (PCA) is widely used as a dimensionality reduction technique. It helps reduce the number of features (i.e., dimensions) by finding, separating out, and sorting the features that explain the most variance in the data in descending order.
Use in *picture classification*: Though PCA does not throw away every other pixel and it only transforms the data to have important features, reducing the dimension to say 100-200 features can be too low. You cannot represent a good image with that.

Datacompression. Other techniques. Manifold learning.



Unroll a “roll”, to obtain a 2d dataset.

https://se.mathworks.com/matlabcentral/fileexchange/36141-laplacian-eigenmap-diffusion-map-manifold-learning

Model optimization. Model building. ML pipeline. Putting it all together.

Example of Machine Learning pipeline.

Exercise 4.

See Canvas.

Putting it all together.
An example.
(Parts of) a machine learning pipeline,
in very, very broad terms.

Model optimization. Model building. ML pipeline. Putting it all together.

Exercise 4.

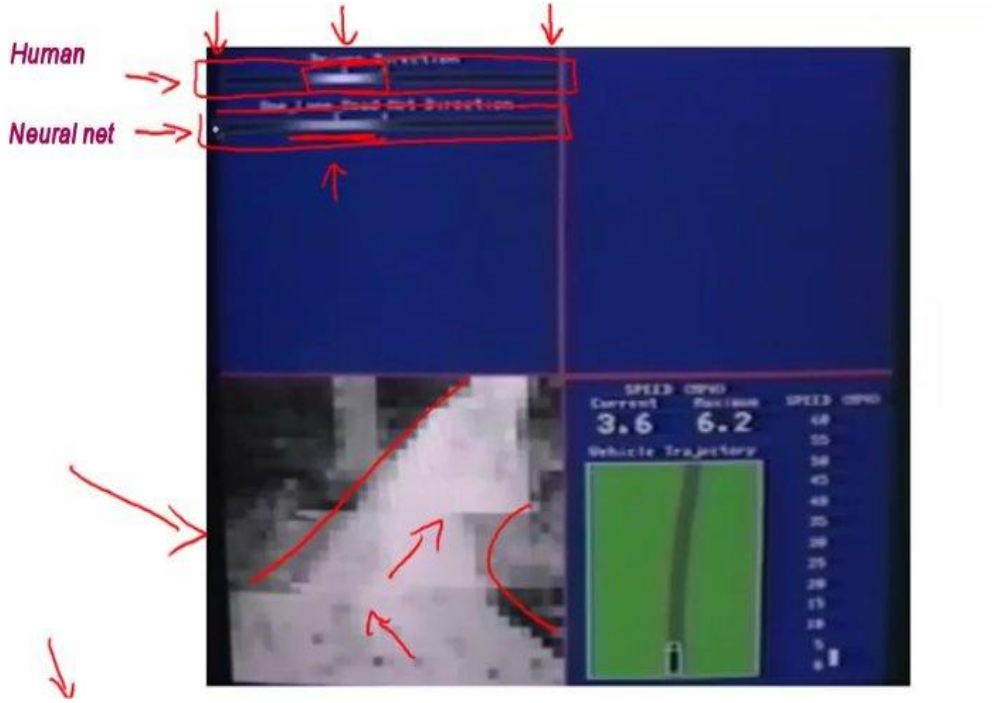
Machine learning pipelines – examples.
Steps to go through (in real life).

In groups of 2-3:
(in very, very broad terms)

- What kind of steps to you think we would need in a machine learning pipeline (given pictures, and ?) to get a car to learn how to stay on the road (say with markings/stripes on the road to indicate the sides of the road).

- Use your imagination: What kind of steps (in very, very broad steps) do you think an ML system should go through to learn the car system to read signs near the road? To spot people on the road?

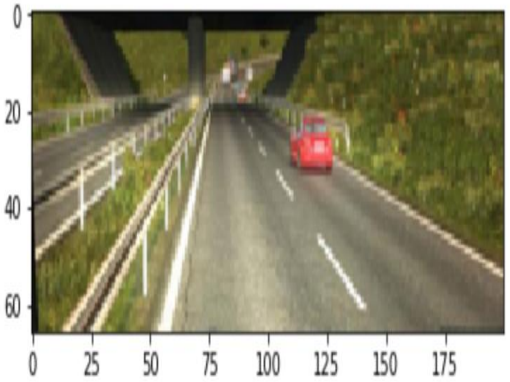
Model optimization. Model building. ML pipeline. Exercise 4 – Discussion.



Staying on the road with a neural net.

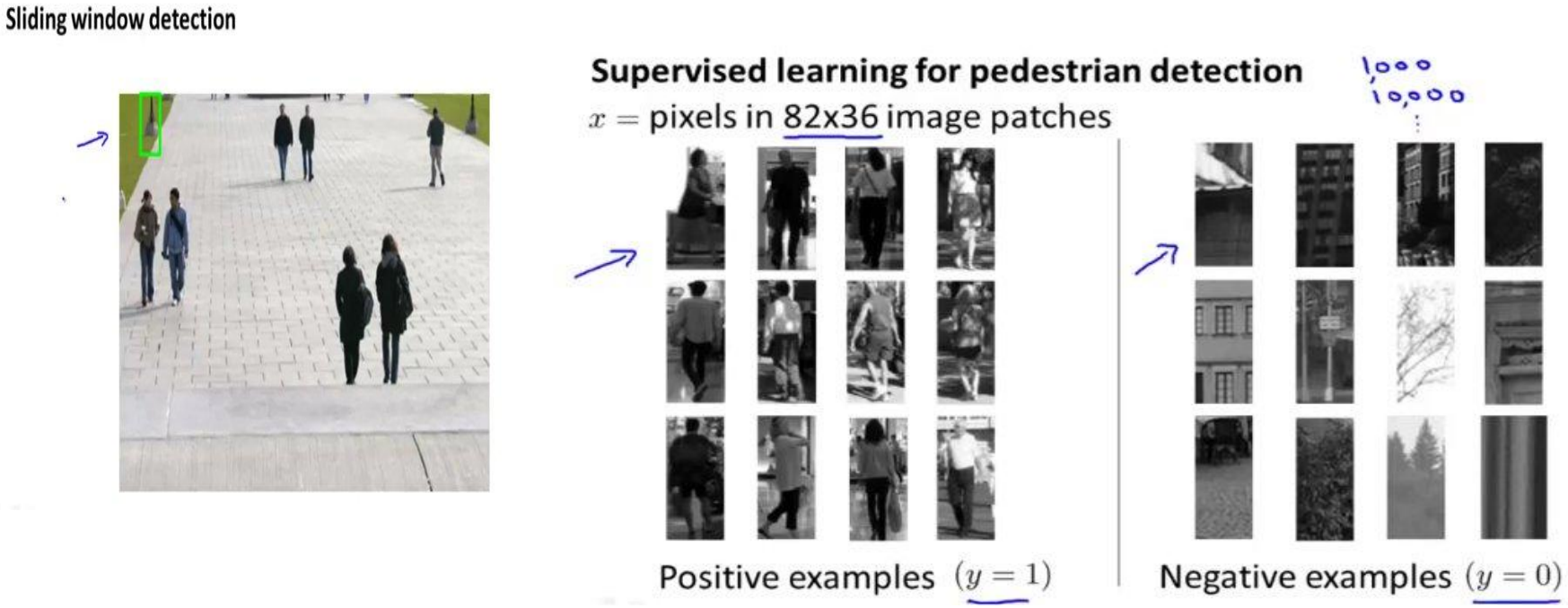
Model optimization. Model building. ML pipeline. Exercise 4 – Discussion.

ground_truth= -81 , prediction= -52.85606846809387



End to End Learning for Self-Driving Cars. Nvidia.

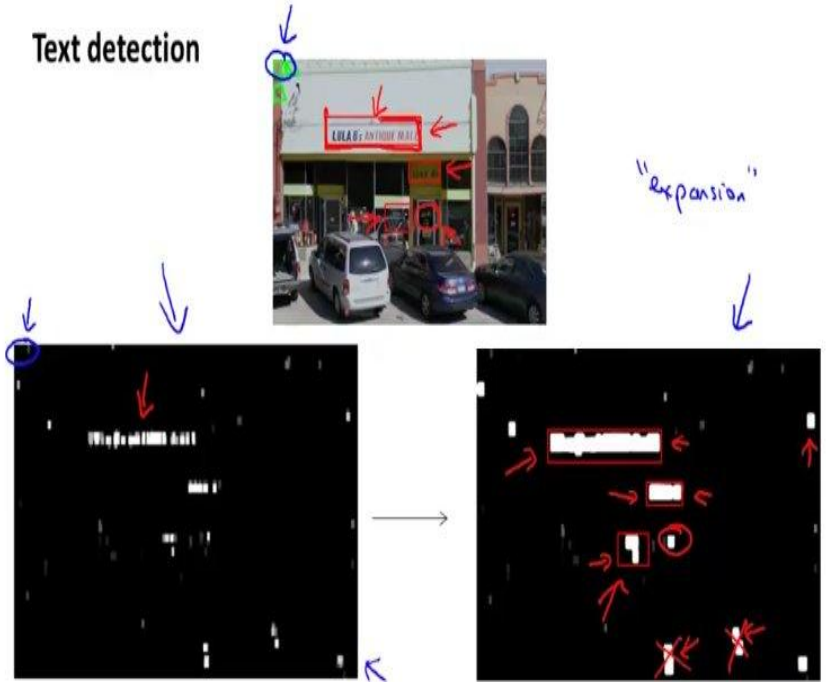
Model optimization. Model building. ML pipeline. Exercise 4 – Discussion.

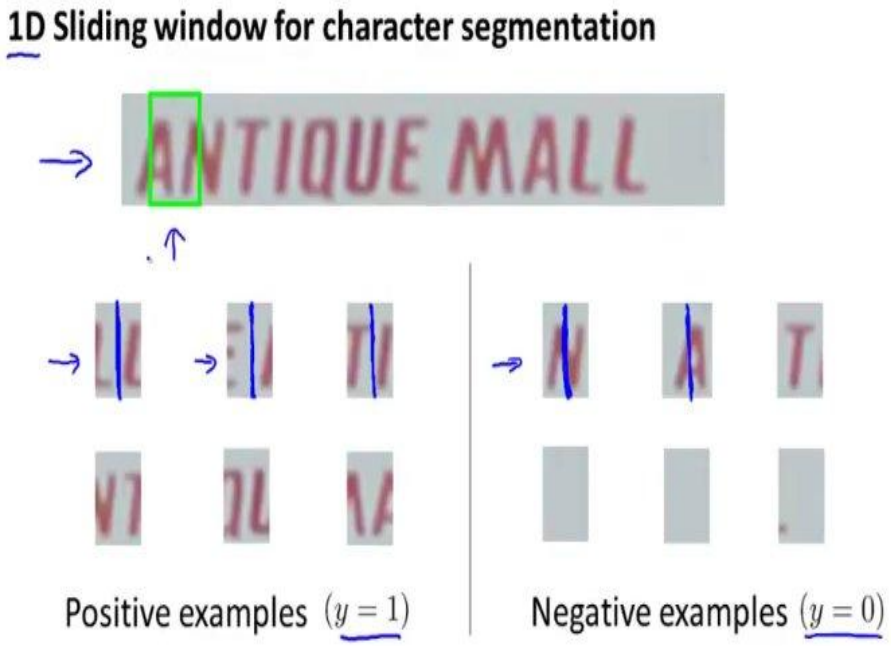


Finding people in an image. Andrew Ng, Coursera.

Model optimization. Model building.

ML pipeline. Exercise 4 – Discussion.





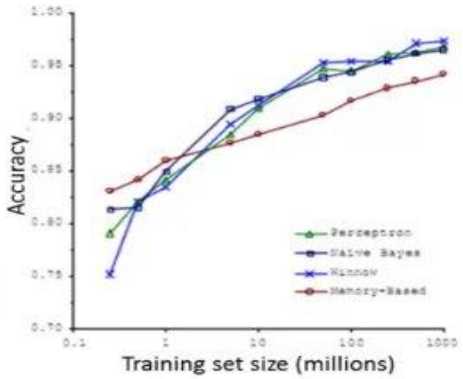
Reading characters in an image. Andrew Ng, Coursera.

Machine Learning – Large datasets.

Machine learning and data

Classify between confusable words.
E.g., {to, two, too}, {then, than}.

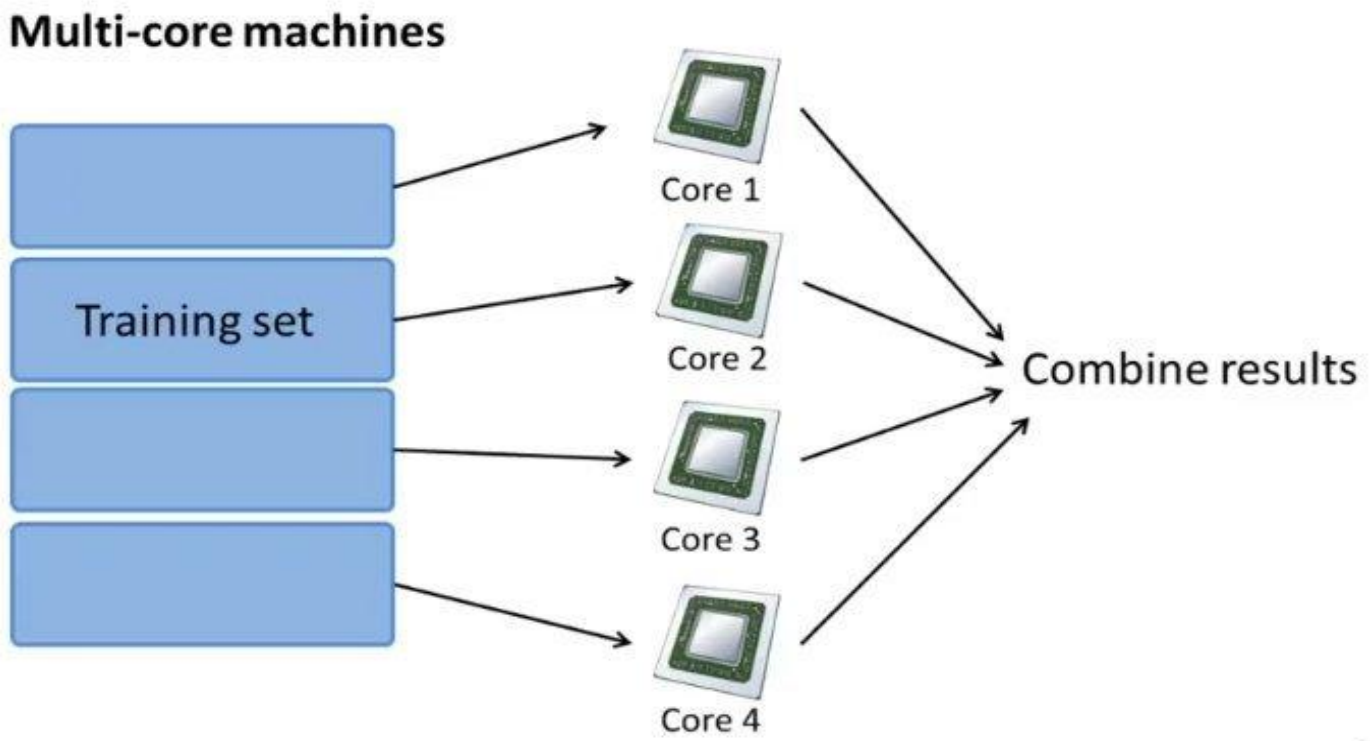
For breakfast I ate two eggs.



→ "It's not who has the best algorithm that wins.
It's who has the most data."

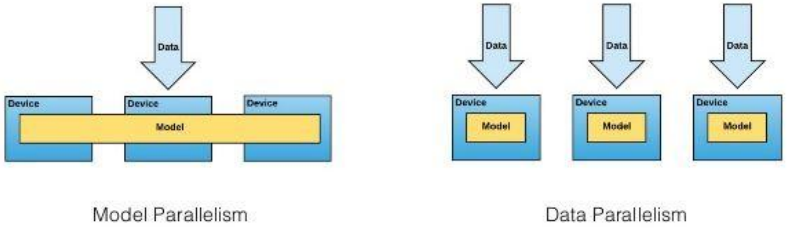
Andrew Ng. Coursera.

Machine Learning – Large datasets.



Use a lot of computer power.....

Machine Learning – Large datasets.



Model Parallelism Data Parallelism

In **data-parallelization** we run the same model over different datastreams, and then synchronize the weights, θ, in the model
(I.e. training different layers in a Deep Learning model in parallel on different GPUs would then be ''**model-parallelism''**).

See: Horovod
https://github.com/horovod/horovod

Machine Learning –
The machine learning pipeline. Putting it all together.

- Start with a simple algorithm that you can implement quickly. Implement it and test it on your cross validation data.

- Plot learning curves to decide if more data, more features are likely to help.

- Error analysis. Manually examine the examples (in the cross validation set) where your model makes errors. Can you spot any systematic errors in the types of examples that your model fails on?

Machine Learning – Unsupervised learning.

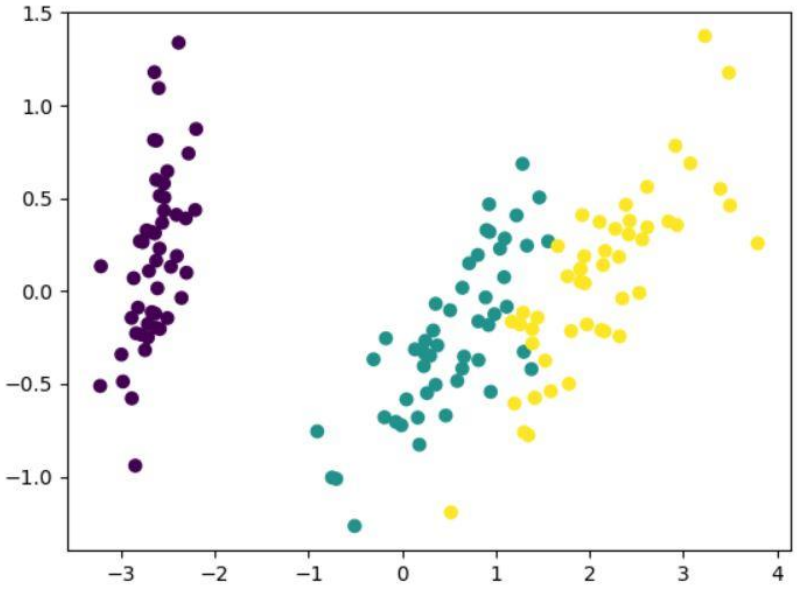
Remaining Exercises...

Exercise 5.

Machine Learning – Unsupervised learning.

Remaining Exercises...

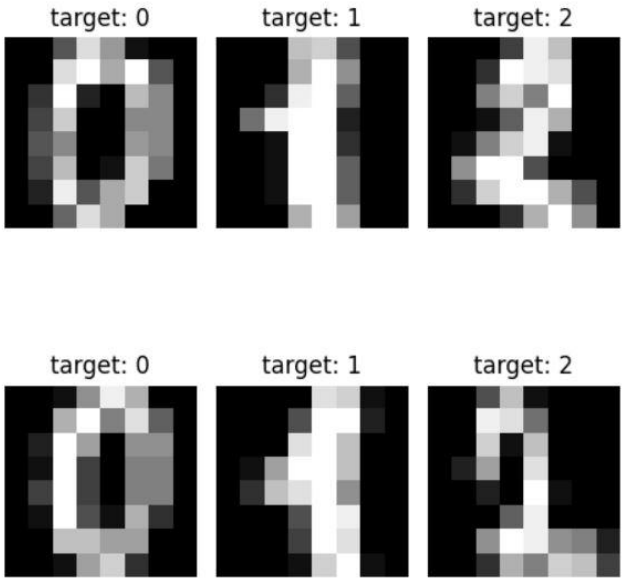
Exercises 6.



Machine Learning – Supervised learning.

Remaining Exercises...

Exercise 7.



Homework

Do the exercises for this week.

Read:
"Hands-On Machine Learning" by Aurelien Geron.
Chapter 3. p. 108 – 118 (v2. p. 90 – 100) (Confusion Matrix).
Chapter 8. 243 – 246 (v2. p. 213 – 224)
(PCA – Dimensionality Reduction).

"Real-World Machine Learning" by Brink et al.
Chapter 4. (Model evaluation and optimization).
p. 77 – 96
Chapter 5. (Basic feature engineering).
p. 106 - 126