cphbusi

COPENHAGEN BUSINESS ACADEMY

Significance tests and deep learning

Jens Egholm Pedersen
<jeep@cphbusiness.dk>

Recap

- Populations and samples
  - Cross-validation
- Linear regression
  - Multivariate regression
  - Polynomial regression
  - Logistic
- Vectors, matrices and tensors
  - Dimensionality and dimension reduction
- Classification
  - Clustering
- TF/IDF
-

Goal of this block

- Have a basic understanding and knowledge of various terms, models and tests in statistics.
- Compute basic statistics on data using the Python's scientific stack and the Sklearn library.
- Develop an informed guess of when to choose a certain model to answer a concrete type of question and apply technology appropriately.

See also: BI plan

cphbusiness

Goal for today

- Inference tests
  - Significance tests
- Perceptrons
- Accuracy, precision, recall and F1 score
- Deep learning
  - Text generation
- Machine learning tips and tricks

See also: BI plan

Hand-in 6

- Good work generally

- Text is important
  - Solving the task is one thing
  - Understanding it is another (see Bloom's taxonomy)

- Explain the numbers
  - Don't just report an accuracy of 0.9
  - What does an accuracy of 0.9 mean?

Inference tests

- What do you think ‘inference test’ mean?

Inference tests

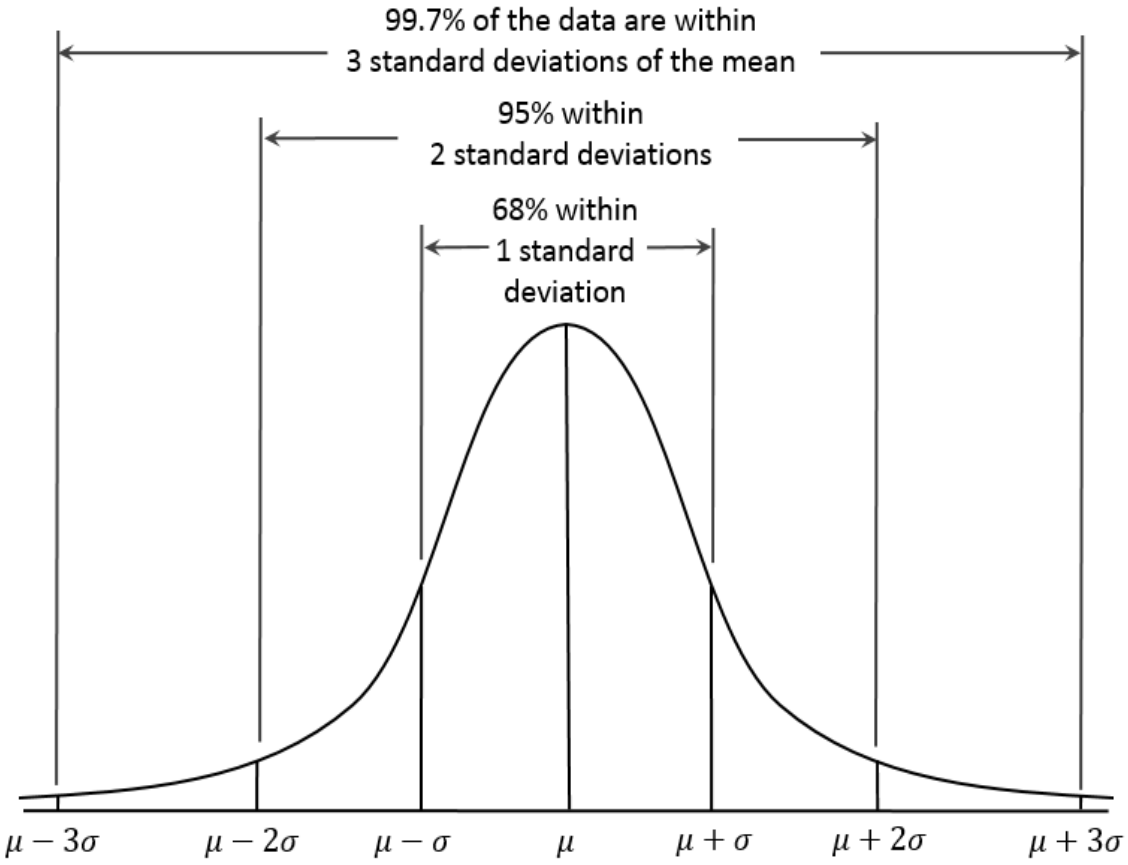
- Testing if we can infer information from a sample

- Mean
- Standard deviation

- How do we know whether we can expect the same to apply for our population?

Probability distribution

- We are looking for the probability that our sample is equal to the population



Probability distribution

- We are looking for the probability that our sample is equal to the population

- A probability distribution over a sample **provides a likelihood that a value would equal that sample**

Hypothesis

- A hypothesis is a statement about a population
  - Usually predicts some parameter value or range of values
  - Example: 68% of the population is above 30 years old

Hypothesis: Men's Corp

- The company *Men's Corp* employs 90% men
- Women are complaining that men are picked more often than women
- What are the population(s) and sample(s)?
- What is the hypothesis?

Hypothesis: Men's Corp

- The company *Men's Corp* employs 90% men
- Women are complaining that men are picked more often than women
- Hypothesis:
  - The company is picking men at a higher rate than in the population
  - Or: Company % of men > population % of men

Hypothesis

- A hypothesis is a statement about a population
  - Usually predicts some parameter value or range of values
  - Example: 68% of the population is above 30 years old
  - Also called the research hypothesis

- Null hypothesis (h_0)
  - A hypothesis stating that something has no effect or are the same
  - Example: the age distribution is the same in two samples
    - The age distribution does not differ i. e. has no effect/difference

Hypothesis: Men's Corp

- The company *Men's Corp* employs 90% men
- Women are complaining that men are picked more often than women
- Research hypothesis:
  - Company % of men > population % of men
- Null hypothesis:
  - Company % of men == population % of men

Hypothesis test

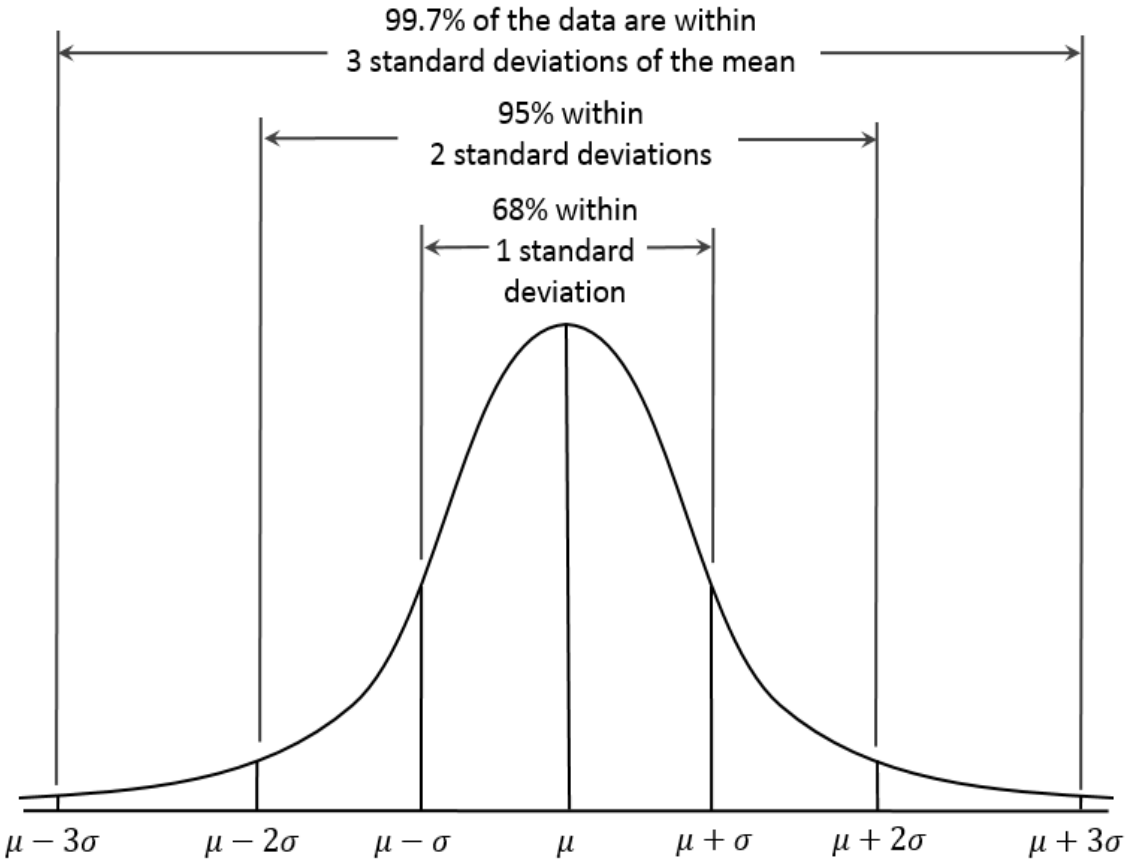
- Research hypothesis
  - What we would like to examine
- Null hypothesis
  - Status quo

- What do we need to prove our research hypothesis?

- We have to show that it is *unlikely* that h_0 is true

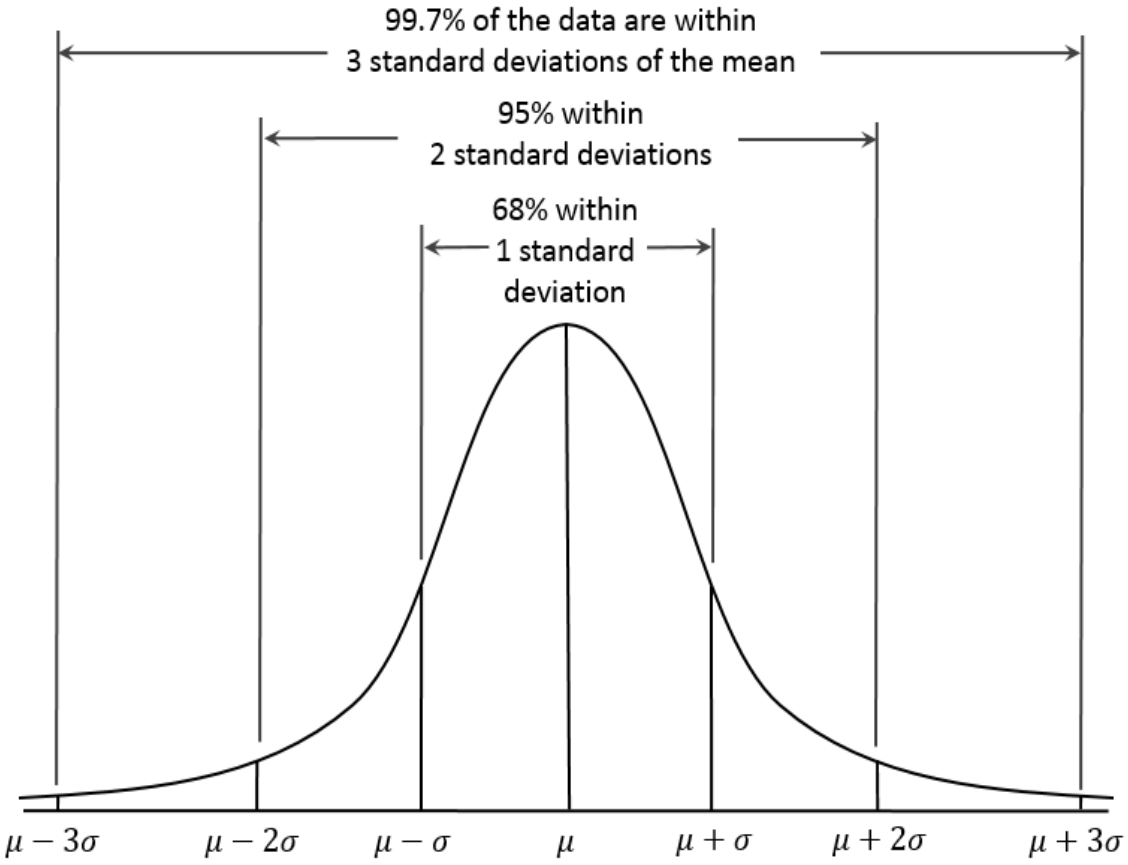
Probability distribution

- We are looking for the probability that our sample is equal to the population



Probability distribution

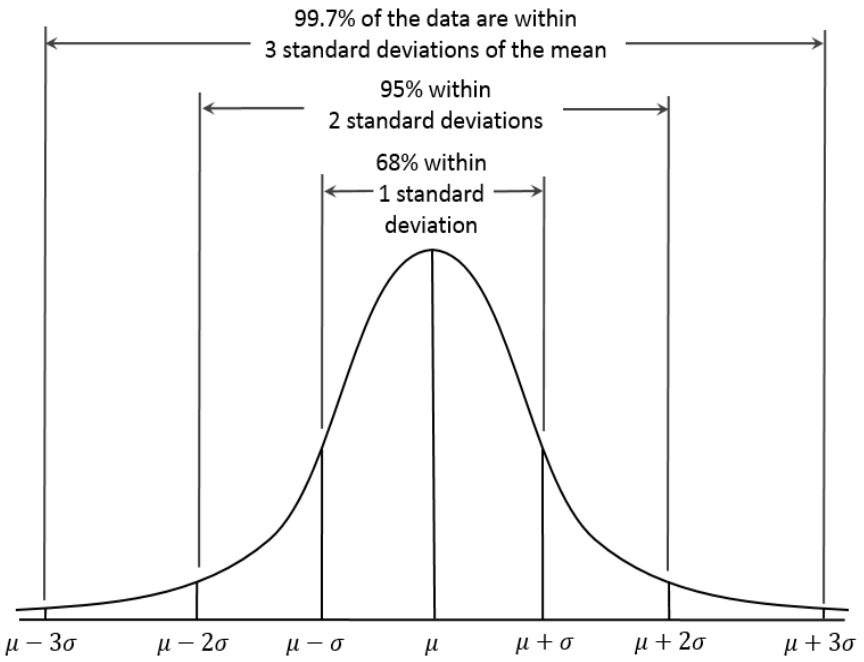
- Also called a sampling distribution
  - Shows the likelihood of a value in the sample



Probability distribution

- Also called a sampling distribution
  - Shows the likelihood of a value in the sample

- Given the population and the sample (the company)
  - How can you express h_0?
  - What about the research hypothesis?



Hypothesis: Men's Corp

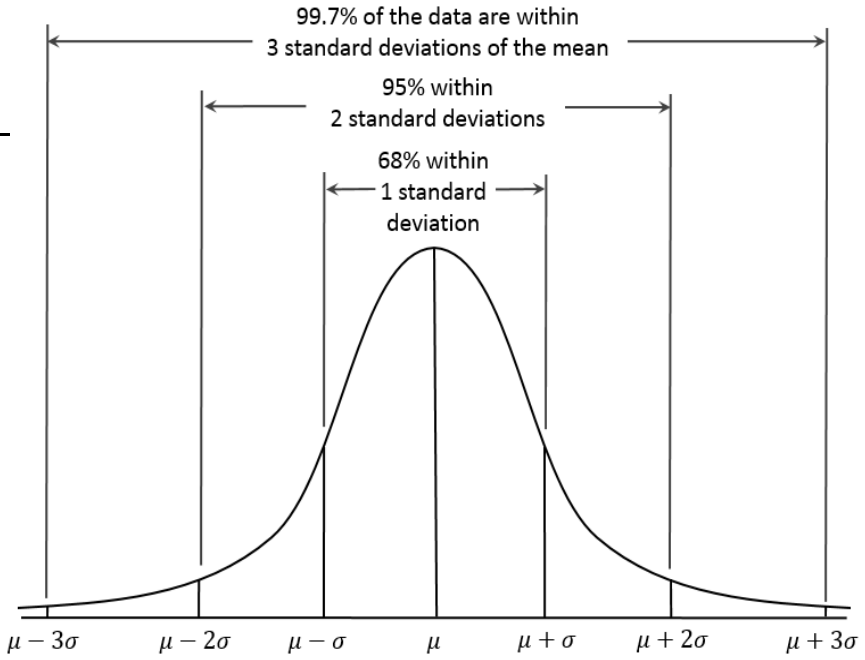
- The company *Men's Corp* employs 90% men
- Women are complaining that men are picked more often than women
- Research hypothesis:
  - Company % of men > population % of men
  - Mean of company men > mean of population men
- Null hypothesis:
  - Company % of men == population % of men
  - Mean of company men == mean of population men

t values

- Test statistic: the difference in the mean between sample and population

$$t = \frac{population\ mean - h_0\ mean}{distribution\ standard\ deviation}$$

$$t = \frac{\mu - \mu_0}{\sigma}$$

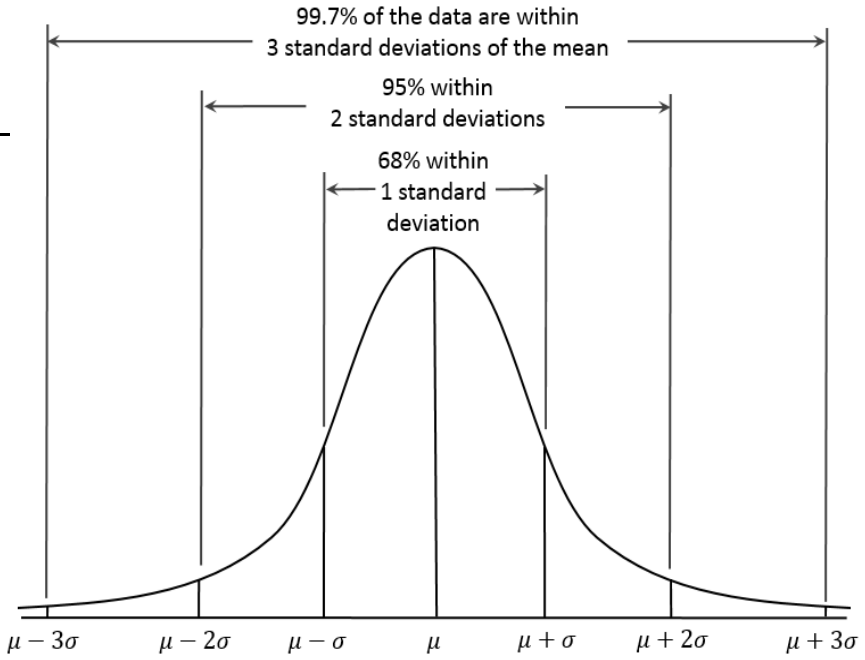


P values

- The *t* value tells us the 'distance' of the means
- Because of the central limit theorem we can convert that to a likelihood!

$$t = \frac{population\ mean - h_0\ mean}{distribution\ standard\ deviation}$$

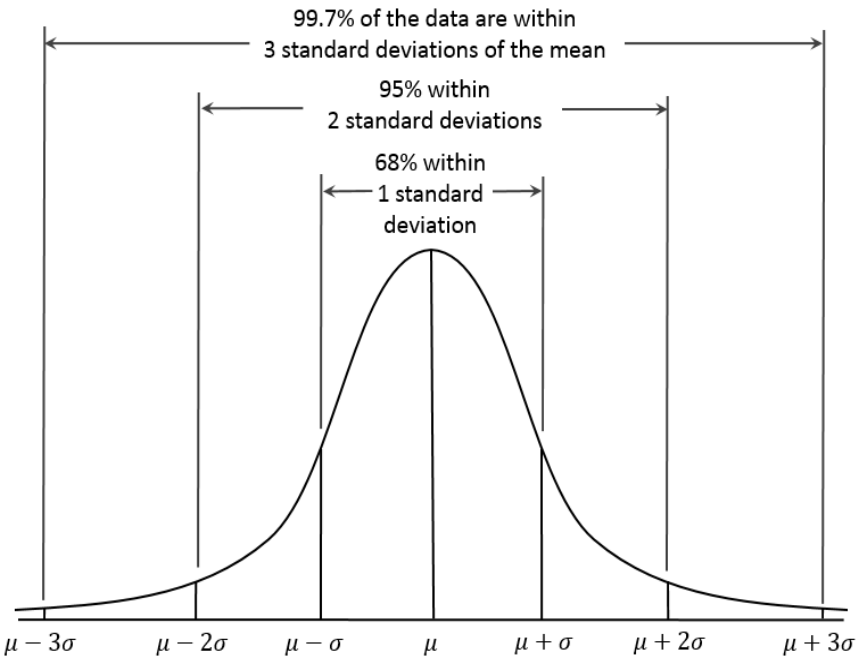
$$t = \frac{\mu - \mu_0}{\sigma}$$



- Example:
  - *t* = 1
  - *t* = 2

P values

- The *t* value tells us the ‘distance’ of the means
- Because of the central limit theorem we can convert that to a likelihood!
- *P* values are deemed **significant** when they are <= 0.05



Significance test / *t* test

- This is called the significance test or *t* test
- We test for the likelihood that we are right
- Testing how likely your sample is according to an expected mean:

```
import scipy.stats as stats
stats.ttest_1samp(arr, 1)
```

Significance test / *t* test

- This is called the significance test or *t* test
- We test for the likelihood that we are right
- Or testing how likely your sample is, related to some other sample:

```
import scipy.stats as stats
stats.ttest_ind(sample1, sample2)
```

Recap

- Inference tests
  - Significance tests
- Perceptrons
- Accuracy, precision, recall and F1 score
- Deep learning
  - Text generation
- Machine learning tips and tricks

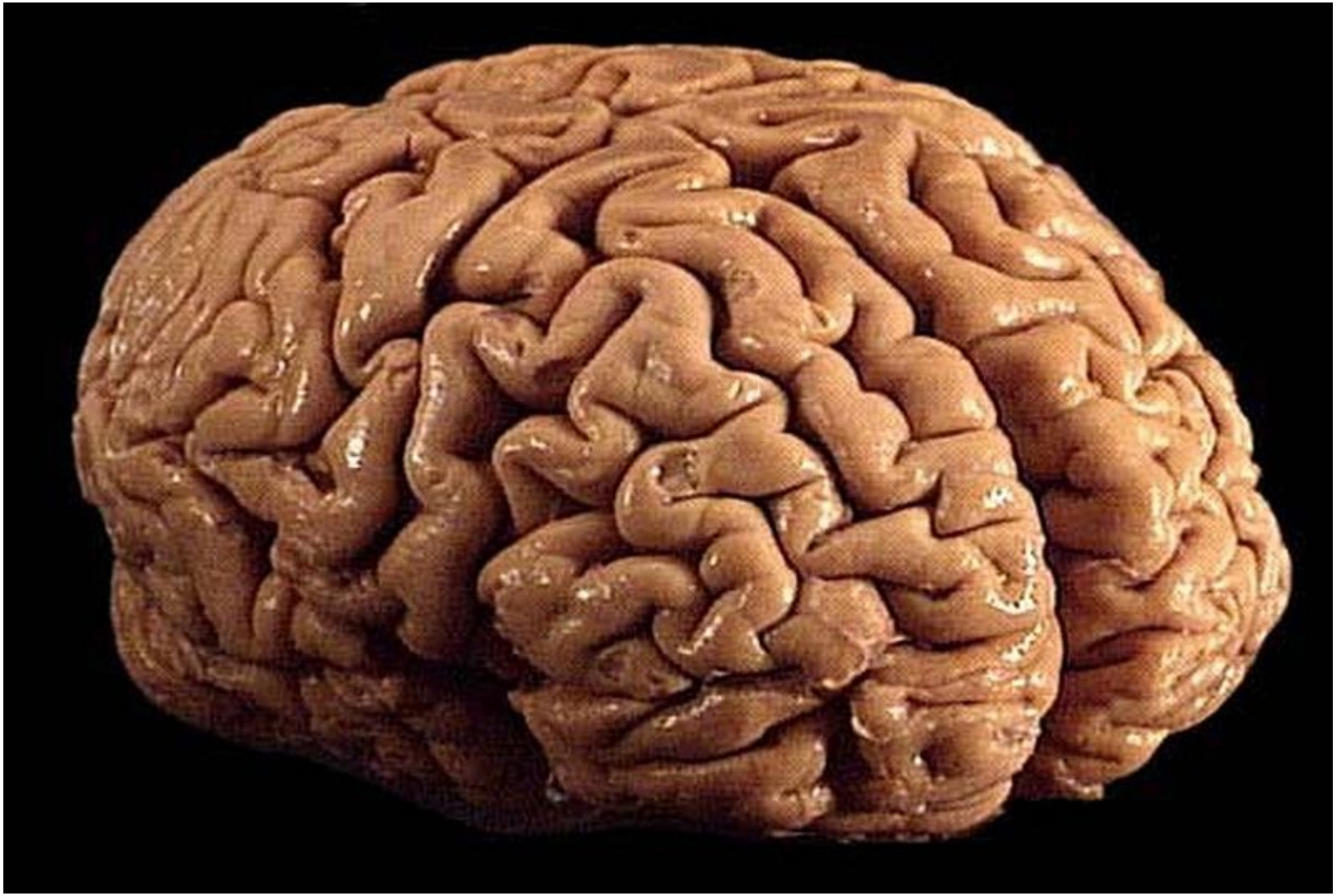
See also: BI plan

Types of machine learning

- Is it trained by a human
  - Supervised / unsupervised
- Can they learn on the fly?
  - Online learning / offline (batch) learning
- Do they include new data?
  - Instance-based / model-based
- Today: supervised, online, model-based learning

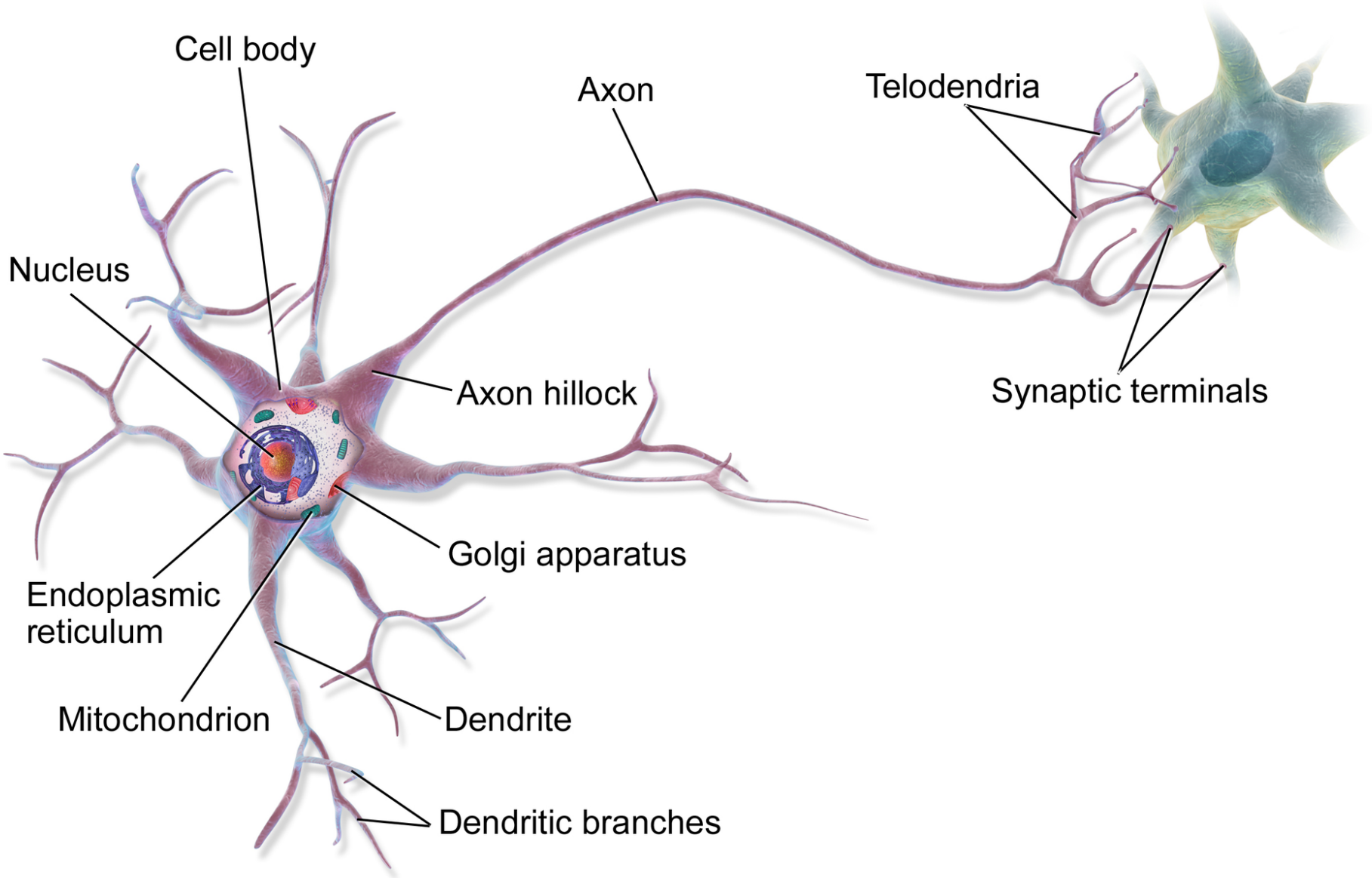
See also: Géron: Hands-on machine learning (book)

The human brain



cphbusiness

The neuron



Perceptron

- Artificial networks is simulating neurons
  - Many inputs, one output

1) Sum all the inputs
2) Does it trigger the neuron?
3) Maybe trigger the output

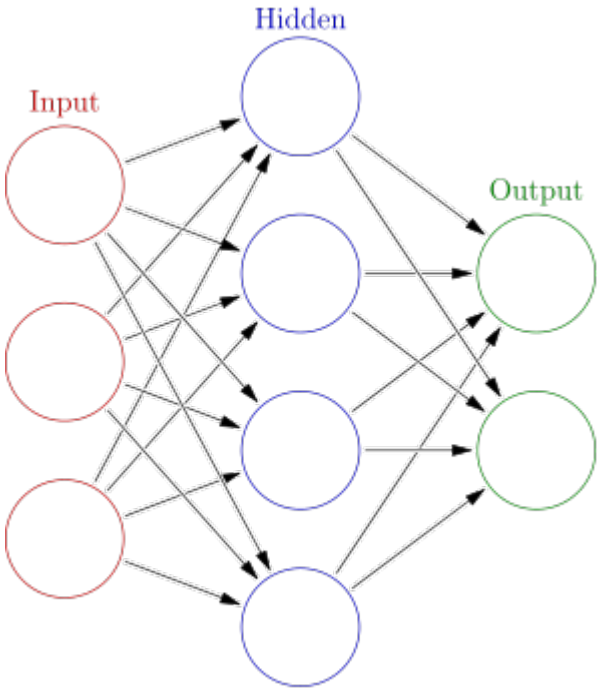
See also: Perceptron on Wikipedia

Perceptron

- Training a neuron
  - Give weights to each input
  - If input triggered the neuron, increment its weight
- Modeling a perceptron
  - Activation function – what makes it ‘fire’
  - Typically a logistic function

Artificial neural networks

- Networks of neurons
  - Typically an input layer and an output layer
  - All layers in between are called hidden layers



Artificial neural networks

- Networks of neurons
  - Typically an input layer and an output layer
  - All layers in between are called hidden layers

- In sklearn

```
from sklearn.linear import Perceptron
model = Perceptron()
model.fit()
```

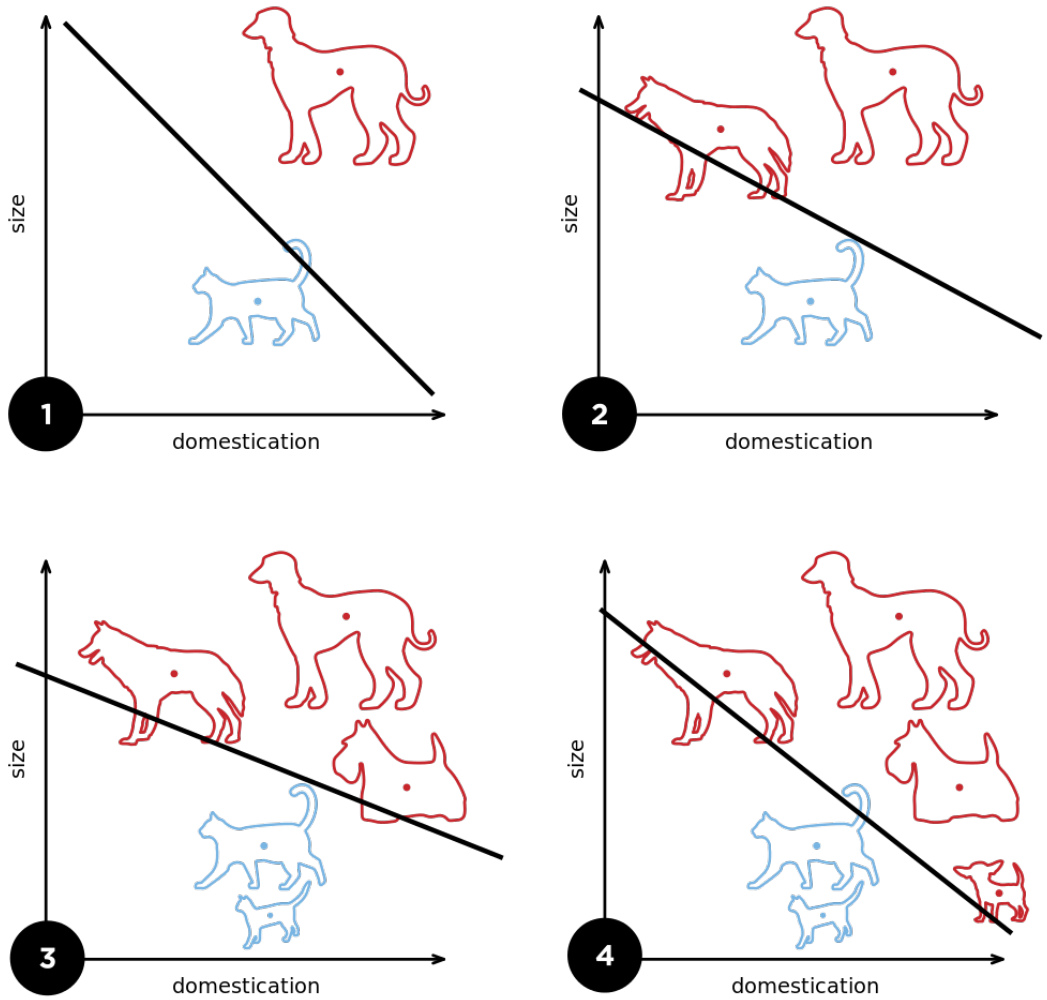
How NN works

- A neuron gets inputs and decides to fire or not
  - Output is always fire / no-fire
  - Classification!

- A neuron is pretty much a linear classifier
  - Yes. This is how you work

cphbusiness

How NN works



See also: Perceptron on Wikipedia, Support Vector Machines (SVM)

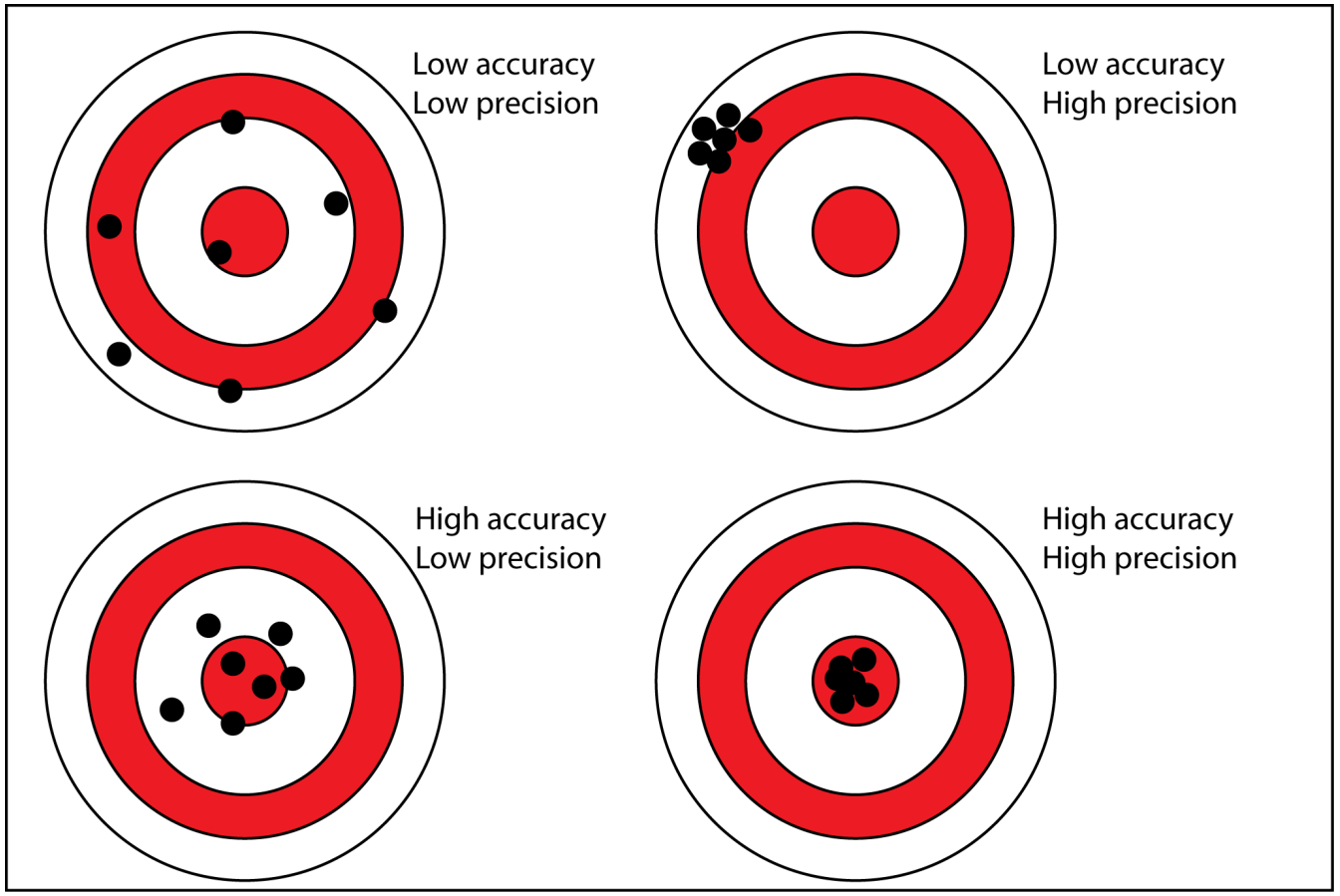
Recap

- Inference tests
  - Significance tests
- Perceptrons
- Accuracy, precision, recall and F1 score
- Deep learning
  - Text generation
- Machine learning tips and tricks

See also: BI plan

Accuracy versus precision

- Why is it not the same?



Confusion matrix

- Accuracy or precision counts correct guesses
  - The number of correct guesses divided by total guesses

- Imagine you are trying to predict cat/non-cat
  - You correctly predict 5 cats out of 27
  - You wrongly predict 2 non-cats as cats out of 27
  - Where is the false and true negatives?!

See also: BI plan

cphbusiness

Confusion matrix

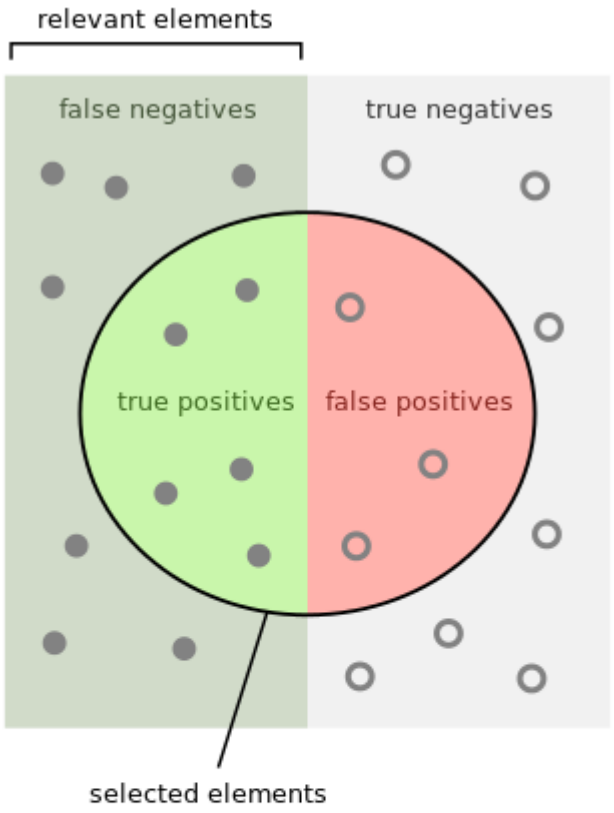
- True or false - positive or negative

		Actual class	
		Cat	**Non-cat**
Predicted class	**Cat**	5 True Positives	2 False Positives
	Non-cat	3 False Negatives	17 True Negatives

`sklearn.metrics.confusion_matrix`

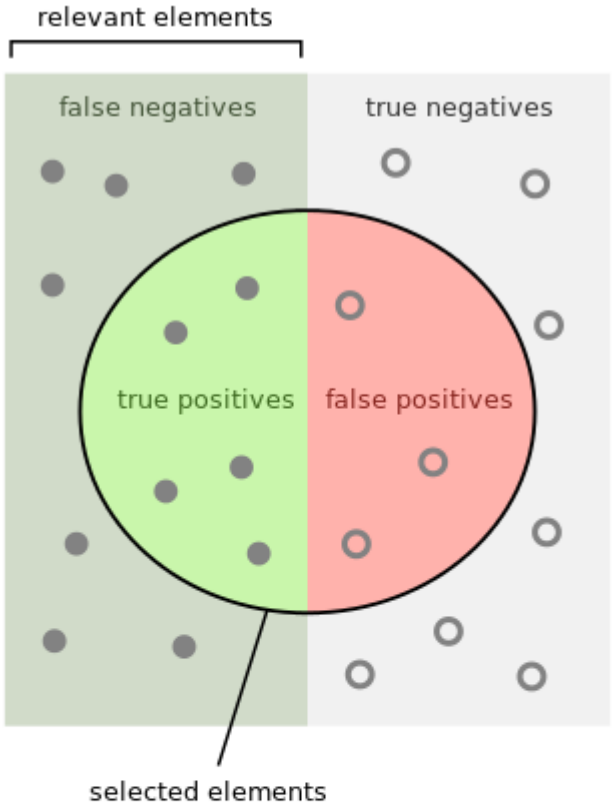
Confusion matrix

- True or false - positive or negative



Precision versus recall

- True or false - positive or negative



How many selected items are relevant?

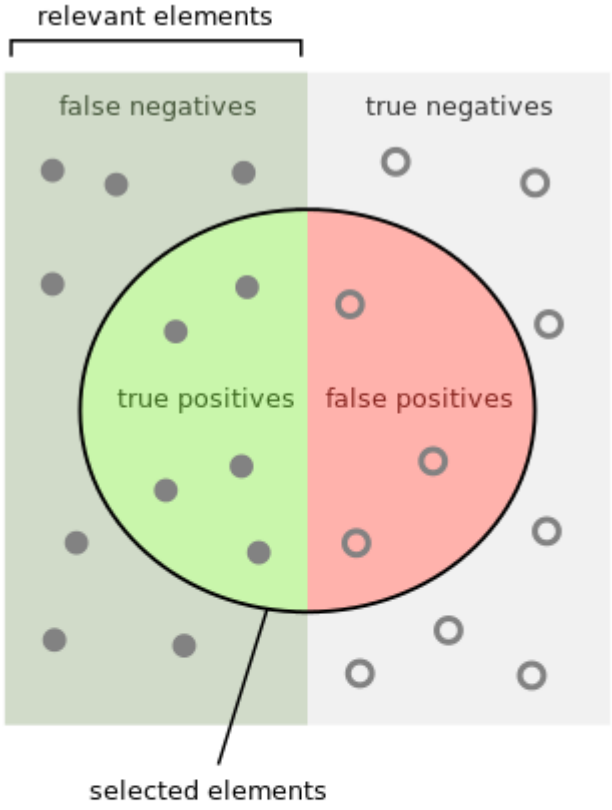
$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

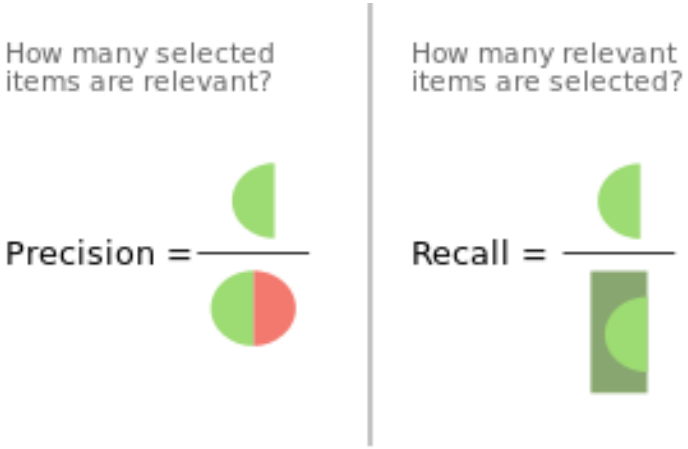
How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

Precision versus recall

- Precision relevant instances among all retrieved
- Recall relevant instances among all relevant





F1 score

- Precision relevant instances among all retrieved
- Recall relevant instances among all relevant

- F1 = 2 * (precision * recall) / (precision + recall)
  - The closer to 1 the better

`sklearn.metrics.f1_score`

Classification reports

- Precision relevant instances among all retrieved
- Recall relevant instances among all relevant
- F1 = 2 * (precision * recall) / (precision + recall)
  - The closer to 1 the better

`sklearn.metrics.classification_report`

Recap

- Inference tests
  - Significance tests
- Perceptrons
- Accuracy, precision, recall and F1 score
- Deep learning
  - Text generation
- Machine learning tips and tricks

See also: BI plan

NN terminology

- Cell
- Weights
- Feedforward versus recurrent networks
- Backpropagation

See also: BI plan

NN problems

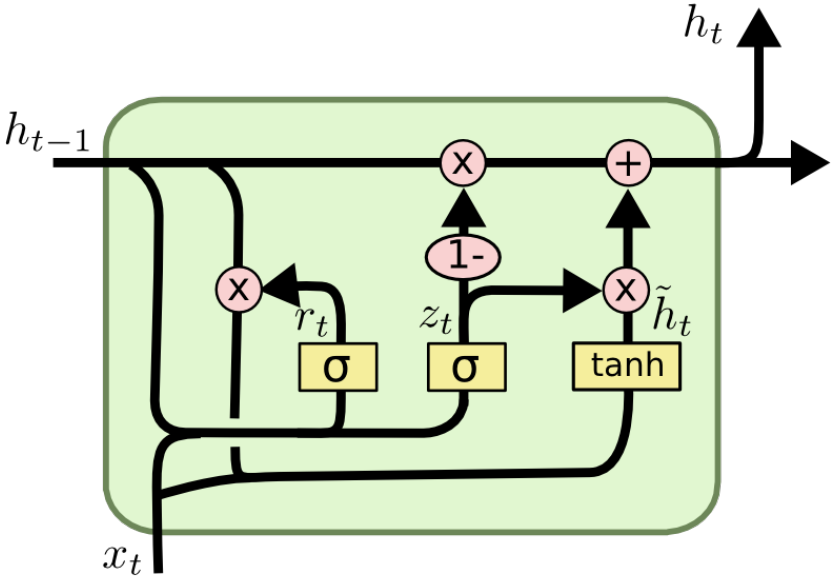
- Problem with NN: fragility and shallowness
- Fragile
  - They “forget” quickly
- They cannot contain “deep” information and complicated representations
  - Example: human visual cortices

Solution: Deep learning

- Problem with NN: fragility and shallowness
- Solution: Deep learning

- Multiple layers
- Multiple levels of abstractions
- Backpropagation

Example: LSTM



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

See also: LSTM on Wikipedia, Understanding LSTM

Tensorflow and Keras

- Tensorflow
  - Tensor = multidimensional (3+) array
  - Library made by Google
- Keras
  - Interface to Tensorflow
  - .. and another ML library called Theano

See also: Keras

cphbusiness

Keras example

- For a single-input model with 2 classes (binary classification):

```
model = Sequential()
model.add(Dense(32, activation='relu', input_dim=100))
model.add(Dense(1, activation='sigmoid'))
model.compile(optimizer='rmsprop',
              loss='binary_crossentropy',
              metrics=['accuracy'])

# Generate dummy data
import numpy as np
data = np.random.random((1000, 100))
labels = np.random.randint(2, size=(1000, 1))

# Train the model, iterating on the data in batches of 32 samples
model.fit(data, labels, epochs=10, batch_size=32)
```

See also: Keras, Keras text generation example

ML resources

- ML cheat sheet
  - https://becominghuman.ai/cheat-sheets-for-ai-neural-networks-machine-learning-deep-learning-big-data-678c51b4b463
- sklearn algorithm cheat-sheet
  - https://cdn-images-1.medium.com/max/1680/1*dYgEs2roROf3j2ANzkDHMA.png
- ML learning repository
  - https://github.com/ageron/handson-ml
  - Also, buy the book!
- Wikipedia on Machine learning
  - Seriously, check the sources

cphbusiness

Next hand-in: Assignment 8

- Deadline: **4th of December 23:59:59**
- Understanding of accuracy, precision and recall
- Understanding populations and t-tests
- Optional part about perceptron network

Next hand-in: Assignment 8

- Deadline: **4th of December 23:59:59**
- The hand-in (on Moodle) should be a link to a GitHub release containing a single file with the code and written text for the assignment parts

- This can either be a .ipynb, .py, .pdf or .md file
- The file must be clearly identifiable. Please name it accordingly (for instance report.pdf).