Computational Communication Science 2 Week 8 - Lecture »A Deep Dive into Supervised Machine Learning«

Marthe Möller
Anne Kroon

a.m.moller@uva.nl, @marthemoller
a.c.kroon@uva.nl, @annekroon

May, 2022

Digital Society Minor, University of Amsterdam

Today

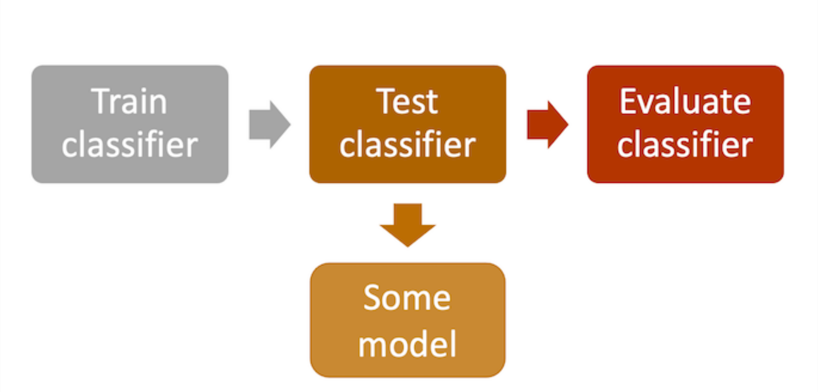
Recap

Recap

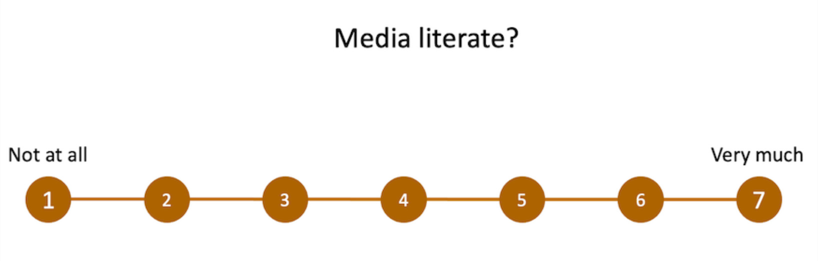
Short recap of last week.

Some classical ML models

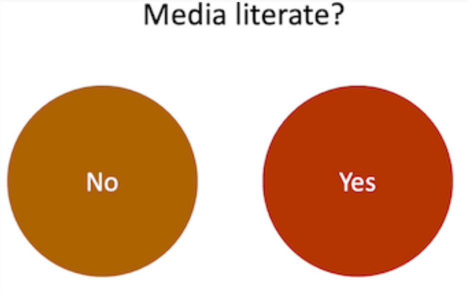
Regression


Train classifier
Test classifier
Evaluate classifier
Some model

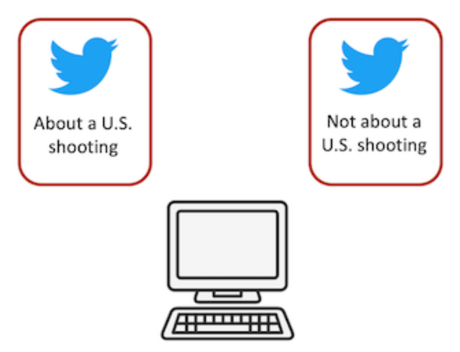
Regression



Regression

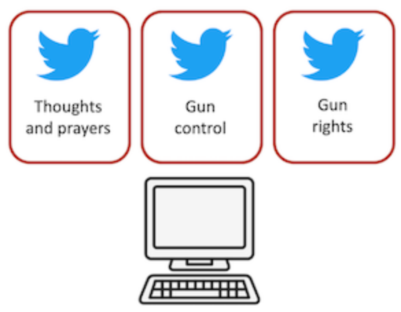


Regression



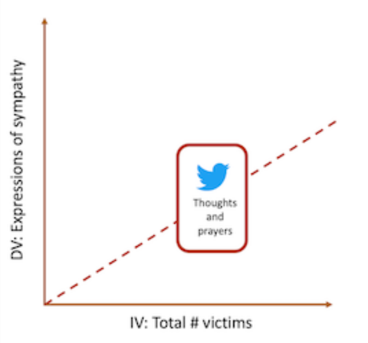
Zhang, Y., Shah, D., Foley, J., Abhishek, A., Lukito, J., Suk, J., Kim, S. J., Sun, Z., Pevehouse, J., & Garlough, C. (2019). Whose lives matter? mass shootings and social media discourses of sympathy and policy, 2012–2014. *Journal of Computer-Mediated Communication*, *24*(4), 182–202. https://doi.org/10.1093/jcmc/zmz009

Regression



Zhang, Y., Shah, D., Foley, J., Abhishek, A., Lukito, J., Suk, J., Kim, S. J., Sun, Z., Pevehouse, J., & Garlough, C. (2019). Whose lives matter? mass shootings and social media discourses of sympathy and policy, 2012–2014. *Journal of Computer-Mediated Communication*, *24*(4), 182–202. https://doi.org/10.1093/jcmc/zmz009

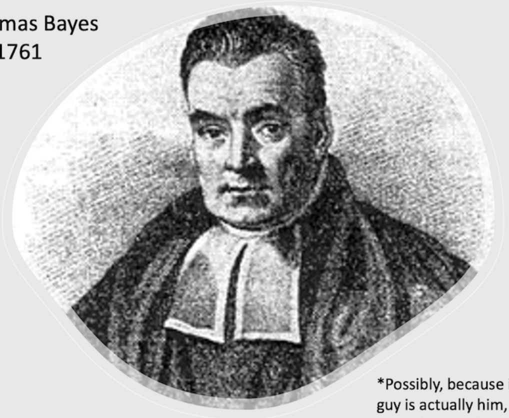
Regression



Zhang, Y., Shah, D., Foley, J., Abhishek, A., Lukito, J., Suk, J., Kim, S. J., Sun, Z., Pevehouse, J., & Garlough, C. (2019). Whose lives matter? mass shootings and social media discourses of sympathy and policy, 2012–2014. *Journal of Computer-Mediated Communication*, *24*(4), 182–202. https://doi.org/10.1093/jcmc/zmz009

Naïve Bayes

Possibly* Thomas Bayes
1702 – 1761



*Possibly, because it is unclear if this guy is actually him, but there is no other (claimed) portrait of him.

Naïve Bayes

$P(\mathrm{A} \mid \mathrm{B}) = \frac{\mathrm{P(B|A) \cdot P(A)}}{\mathrm{P(B)}}$

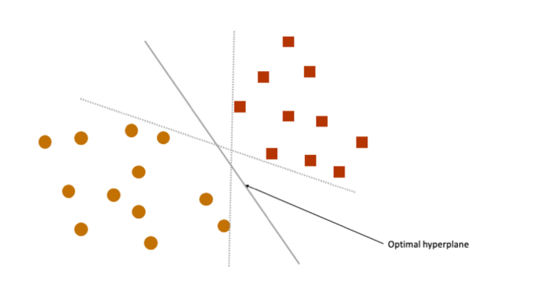
Mathematicians' language for: the probability of A is B is the case/present/true.

$P(\mathrm{label} \mid \mathrm{features}) = \frac{\mathrm{P(features|label) \cdot P(label)}}{\mathrm{P(features)}}$

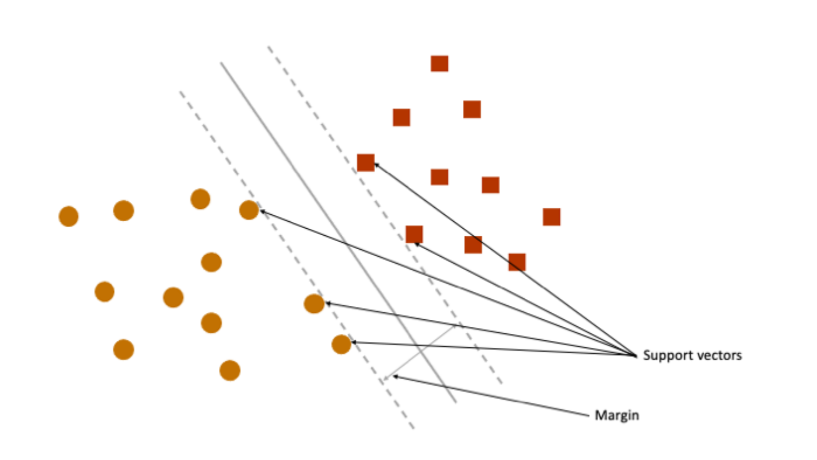
Support Vector Machines

SVMs aim to find a hyperplane in an N-dimensional pace that distinctly classifies the datapoints.

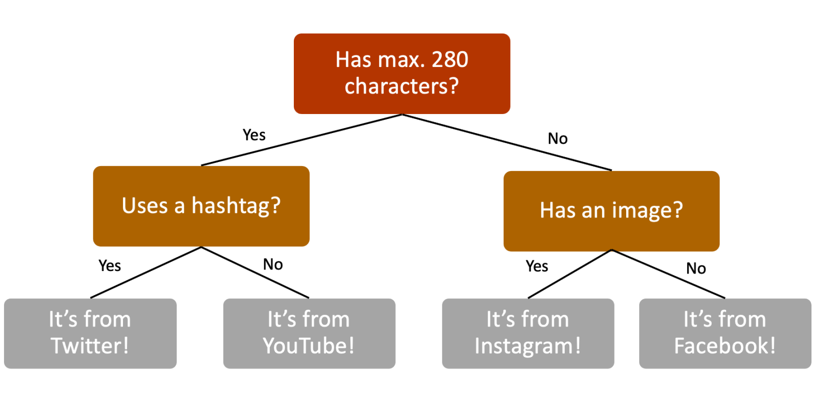
The best hyperplane is the one that has the maximum margin (distance) between the datapoints of both classes.



Support Vector Machines



Decision Trees and Random Forests



Decision Trees and Random Forests

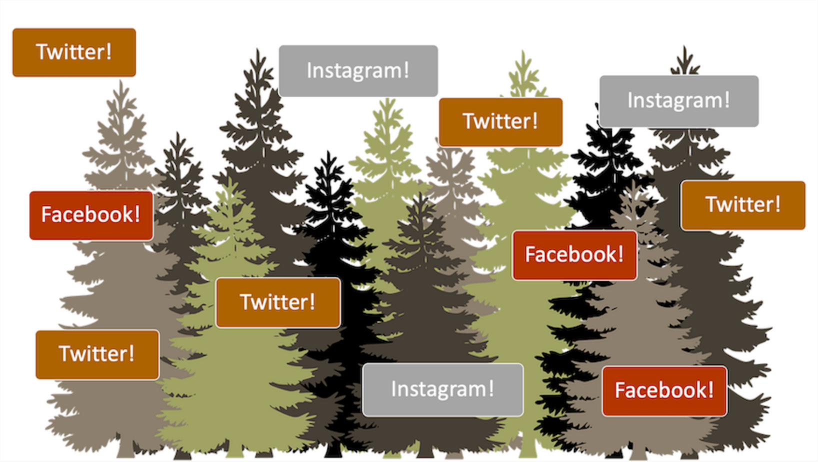
Advantages of decision trees:

- Transparency
- Suitable for non-linear relationships

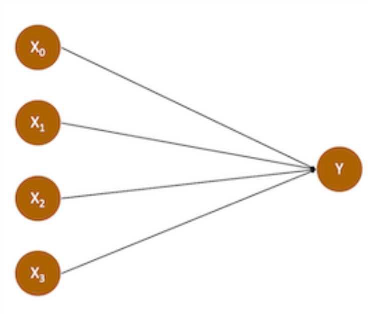
Disadvatanges of decision trees:

- Loss of nuance due to yes/no-design
- Cannot correct early mistakes
- Prone to overfitting

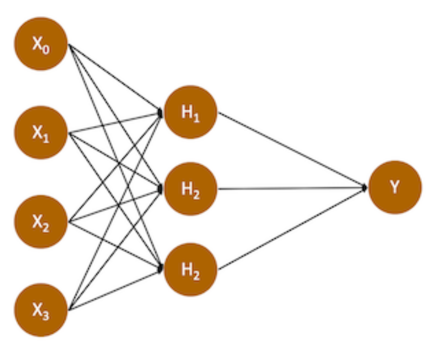
Decision Trees and Random Forests



Neural Networks



Neural Networks



Ha, Y., Park, K., Kim, S. J., Joo, J., & Cha, M. (2021). Automatically detecting image–text mismatch on instagram with deep learning. *Journal of Advertising*, *50*(1), 52–62. https://doi.org/10.1080/00913367.2020.1843091

Recap

Many different models available for machine learning.

How do you know what is the best for your case? Try it out and validate!

Zooming out

We talked about:

- The principles behind SML
- The steps of SML
- Some commonly used ML models

Next, we will talk about:

- Validating models

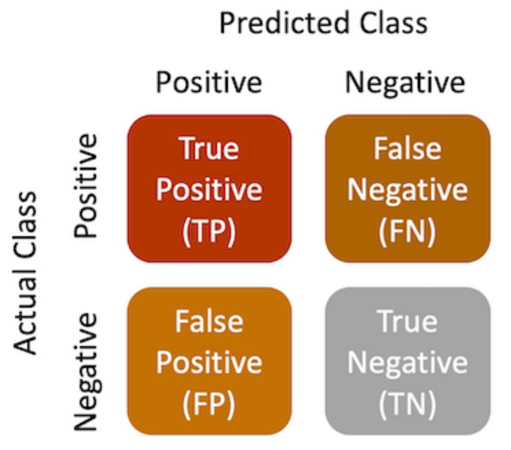
Validating models

Precision and Recall

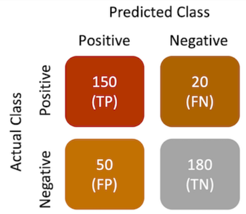
Precision quantifies the number of positive class predictions that actually belong to the positive cases.
OR: How much of what we found is actually correct?

Recall quantifies the number of positive class prediction made out of all positive examples in the dataset.
OR: How many of the cases that we wanted to find did we actually find?

Precision and Recall



Precision and Recall



Precision is calculated as: $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$

In our case $\frac{150}{150+50}$ which is 0.75

Recall is calculated as $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$

In our case $\frac{150}{150+20}$ which is 0.88

Precision and Recall

Table 2
Relationship classification performance and number of training tweets, random sampling approach.

		100	200	500	1000	2000	3000	4000
Linear support vector machine classifier	AC	0.63	0.65	0.70	0.73	0.80	0.84	0.91
	PC	0.45	0.48	0.59	0.62	0.76	0.80	0.90
	RC	0.38	0.43	0.51	0.59	0.71	0.79	0.86
	AUC	0.41	0.45	0.59	0.61	0.69	0.76	0.85
	KA	0.09	0.10	0.39	0.41	0.54	0.65	0.79
Naïve Bayes classifier	AC	0.63	0.65	0.71	0.75	0.82	0.86	0.91
	PC	0.42	0.46	0.62	0.68	0.81	0.86	0.92
	RC	0.27	0.33	0.47	0.49	0.61	0.69	0.79
	AUC	0.33	0.38	0.60	0.62	0.69	0.77	0.84
	KA	0.08	0.13	0.39	0.40	0.56	0.67	0.78
Logistic regression classifier	AC	0.66	0.67	0.71	0.74	0.79	0.85	0.89
	PC	0.48	0.51	0.63	0.70	0.78	0.89	0.93
	RC	0.04	0.22	0.35	0.39	0.53	0.64	0.73
	AUC	0.08	0.31	0.51	0.55	0.62	0.74	0.82
	KA	0.01	0.09	0.21	0.32	0.48	0.64	0.74

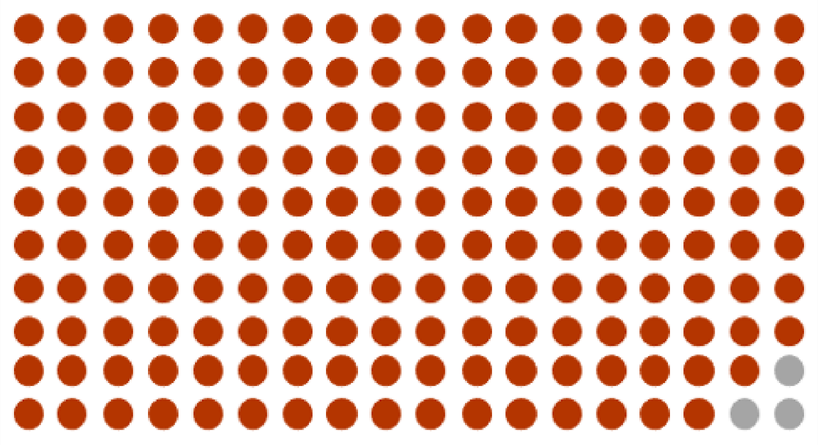
Van Zoonen, W., & Van der Meer, T. G. (2016). Social media research: The application of supervised machine learning in organizational communication research.. *Computers in Human Behavior*, *63*, 132–141.

https://doi.org/10.1016/j.chb.2016.05.028

Accuracy

Accuracy: In which percentage of all cases was our classifier right?
Class distribution: The number of examples that belong to each class.
Imbalanced classification: A classification predictive modeling problem where the distribution of examples across the classes within a training dataset is not equal.

Accuracy



Majority class (red dots) vs. minority class (grey dots)

F_1-score

F_1-score: The harmonic mean of precision and recall.

$F_1\text{-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

Validating Models

Many more metrics to validate models.

Learn more using, for example, the scikit-learn documentation.

Cross-validation

Overfitting: When a model fits *exactly* against the data.

When we calculate the metrics discussed above for multiple models on the same test dataset, we run the risk of overfitting on the test data.

Potential solution: Split the dataset into three smaller sets. A training dataset, a validation dataset and a test dataset.

However, this requires us to have a very large labeled dataset. In reality, this is not always the case!

Cross-validation

Cross-validation: A resampling procedure to evaluate ML models on a limited data sample.

k-fold cross-validation, where k refers to the number of groups or folds in which a sample is split.

k-fold cross-validation

k-fold cross-validation step by step:

1. Shuffle the data
2. Split the data into k folds (groups)
3. For each unique group
   - 3.1 Take the group as a test dataset
   - 3.2 Take the remaining groups as one training dataset
   - 3.3 Fit a model on the training set and evaluate it on the test set
   - 3.4 Retain the evaluation score and discard the model
4. Summarize the evaluation scores to assess the model

Cross-validation

Cross-validation is often used to compare many different model specifications, for example to find the best hyperparameters.

Hyperparameters: Parameters of the model that are not estimated from the data.

To do this, the Grid Search algorithm is often used.

More about hyperparameters in this week's tutorial!

Zooming out

We talked about:

- The principles behind SML
- The steps of SML
- Some commonly used ML models
- Validating models

Next, we will talk about:

- Strengths and challenges associated to SML

SML: Strenghts and Challenges

Strengths and Challenges

Strengths:

- Easier to code large datasets
- Enhances replicable research
- Easier to study ”natural” human behavior

Disadvantages:

- Resource constraints
- Ethical considerations
- Criticism required (see next slide)

Strengths and Challenges

23 november 2021 00:47
Laatste update: 1 dag geleden

761 NUjij-reacties

Het systeem van de Belastingdienst koos ervoor om de kinderopvangtoeslag vooral bij mensen met een laag inkomen extra te controleren. Dat heeft de fiscus toegegeven in antwoord op vragen van *Trouw* en *RTL Nieuws*.

Trouw

Toeslagen

Belastingdienst ging vooral achter lage inkomens aan

Om toeslagen te controleren op fouten en fraude gebruikte de Belastingdienst een zelflerend algoritme. Dat selecteerde vooral lage inkomens voor controle.

Jan Kleinnijenhuis 22 november 2021

De Belastingdienst heeft jarenlang specifiek burgers met een laag inkomen geselecteerd voor extra controle op

Fraudejacht bij Toeslagen

Belastingdienst controleerde extra bij lage inkomens in jacht op fraude

22 november 2021 22:59

Zooming out I

We talked about:

- The principles behind SML
- The steps of SML
- Some commonly used ML models
- Validating models
- Strengths and challenges associated to SML

Zooming out II

This week's tutorial:

- Hands-on approach to take a further look into the machine learning process
- Tutorial goals:
  - To get you some experience with the SML process and selecting a model
  - To provide a stepping stone so that you can (independently) advance your machine learning skills