Supervised machine learning for text classification

Mate Akos

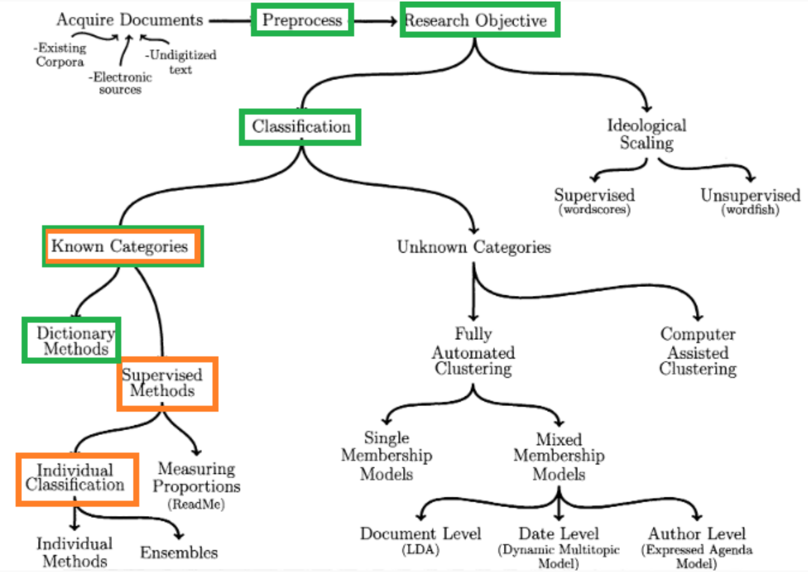
2019 October

TK PTI

For today

1. Supervised vs. unsupervised methods
2. Key concepts
3. Naive Bayes
4. SVM

Where we are



Acquire Documents
-Existing Corpora
-Electronic sources
-Undigitized text
Preprocess
Research Objective
Classification
Ideological Scaling
Supervised (wordscores)
Unsupervised (wordfish)
Known Categories
Unknown Categories
Dictionary Methods
Supervised Methods
Individual Classification
Measuring Proportions (ReadMe)
Individual Methods
Ensembles
Fully Automated Clustering
Computer Assisted Clustering
Single Membership Models
Mixed Membership Models
Document Level (LDA)
Date Level (Dynamic Multitopic Model)
Author Level (Expressed Agenda Model)

Classification

- **Objective**: classify our data into *n* categories.
- Our **response variable** is categorical/qualitative (eg.: gender, pass/fail, etc.)

Key general terms in the statistical learning domain

- X_n: input variables, also called: predictors, independent variables, features
- Y: output variable, also called: response, target or dependent variable

A classification problem

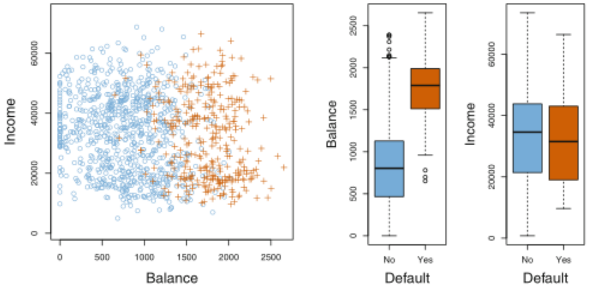


FIGURE 4.1. *The* `Default` *data set.* Left: *The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.* Center: *Boxplots of* `balance` *as a function of* `default` *status.* Right: *Boxplots of* `income` *as a function of* `default` *status.*

James, Witten, Hastie and Tibshirani (2017), p.129

Supervised vs. unsupervised learning

Supervised learning:

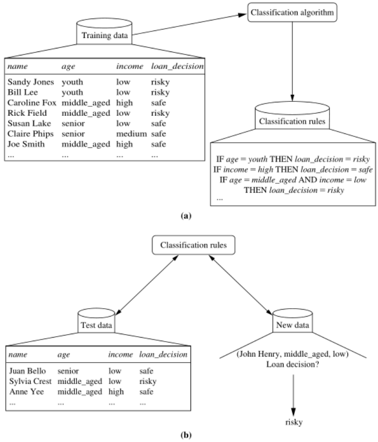
- **Classification problems** where we know the exact categories of our response variable
- We have labelled data, which ”supervises” the training of the classifier
- **Having labelled data is a must!**
- Key terms:
    - **learning algorithm/classifier:** maps documents to classes (in case of text)
    - **training set:** a subset of our labelled data, which is used to train the classifier
    - **test set:** a smaller subset of our data to see how our classifier performs on ’unseen’ data

Supervised vs. unsupervised learning

Unsupervised learning:

- **Clustering problems**, we don’t know the number of categories of our response variable
- The clustering algorithm finds clusters in the data without any ”supervision”
- We don’t have any labelled data
- Useful in exploratory analysis or discovering clusters in our data

A classification problem



8.1 The data classification process: (a) *Learning*: Training data are analyzed by a classification algorithm. Here, the class label attribute is *loan_decision*, and the learned model or classifier is represented in the form of classification rules. (b) *Classification*: Test data are used to estimate the accuracy of the classification rules. If the accuracy is considered acceptable, the rules can be applied to the classification of new data tuples.

Han, Kamber and Pei (2012), p.329

A classification problem

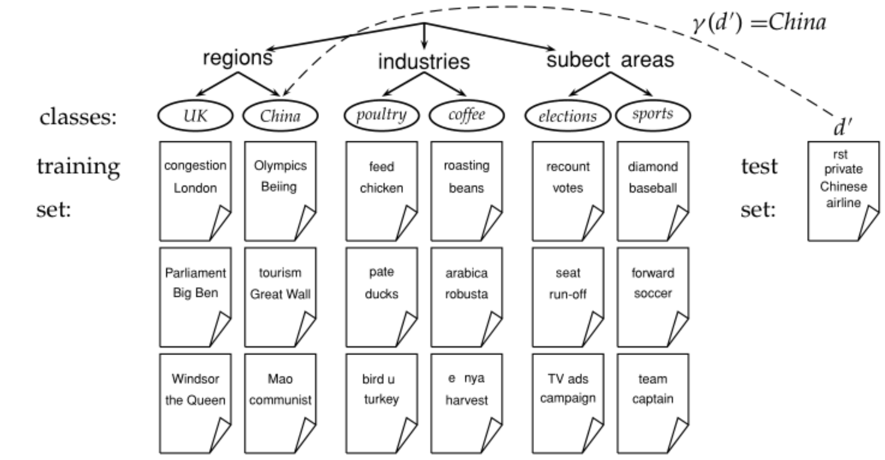


Figure 13.1 Classes, training set, and test set in text classification.

Manning, Raghavan and Schütze (2008), p.238

Asessing classifier performance: confusion matrix

		Predicted	
		cat1	cat0
Actual	cat1	TP	FN
	cat0	FP	TN

- TP for words that are class 1 and predicted in class 1
- FN for words that are class 1 and predicted in class 0
- FP for words that are class 0 and predicted in class 1
- TN for words that are class 0 and predicted in class 0

Metrics from the confusion matrix

- Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
- Precision: $(p) = \frac{TP}{TP+FP}$
- Recall: $(r) = \frac{TP}{TP+FN}$
- Error rate: $ER = \frac{FN+FP}{TP+TN+FP+FN}$
- Specificity = $\frac{TN}{TN+FP}$

Be mindful of the trade-offs. If we want to increase recall by increasing the TP, this will likely also increase our FP, thus lowering specificity.

Also known as:

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

James, Witten, Hastie and Tibshirani (2017), p.149

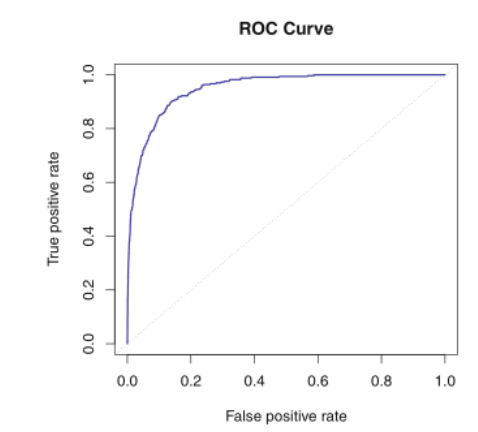
		Predicted	
		cat1	cat0
Actual	cat1	150	40
	cat0	60	250

Accuracy = 0.8

		Predicted	
		cat1	cat0
Actual	cat1	250	45
	cat0	5	200

Accuracy = 0.9

The Receiver Operating Characteristic (ROC) curve



James, Witten, Hastie and Tibshirani (2017), p.148

The Receiver Operating Characteristic (ROC) curve

- **True positive rate** = Recall (or Sensitivity)
- **False positive rate** = 1-Specificity
- Diagonal: random guess
- **(TP, FP)**:
    - (0,0): everything is negative
    - (1,1): everything is positive
    - (1,0): ideal, what we want
- Area under the curve (AOC): ideal: 1, random guess: 0.5

Naive Bayes (NB)

- Classifies documents into categories based on posterior probability
- The posterior is established via the Bayes theorem
- Widely used classifier for texts
- ”naive” because it assumes:
    - conditional independence between words
    - positional independence of words (due to bag-of-words approach)

The Bayes theorem:

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

Where:

- $P(A|B)$ posterior probability (conditional probability of A, given that B occured)
- $P(A)$ prior probability (how likely is event A, without any additional information)
- $P(B|A)$ likelihood of B, given A

Naive Bayes

At the term levels, with N documents and K classes:

$$P(c_k|w_j) = \frac{P(w_j|c_k)P(c_k)}{P(w_j)}$$

At the document level

The document level class:

$$P(c|d) = P(c) \prod_j \frac{P(w_j|c)}{P(w_j)}$$

Naive Bayes example

	#	features	is WtP?
training set	1	Forest, Tiger, Forest	yes
	2	Forest, Honey, Forest	yes
	3	Mad, Hatter, Forest	no
	4	Forest, Piglet	yes
test set	5	Forest, Forest, Forest, Mad, Hatter	?

Adapted from Manning, Raghavan and Schütze, Introduction to Information Retrieval (Table 13.1)

NB excercise

- Prior Winnie the Pooh:$P(c) = 3/4$ (for not Winnie: 1/4)
- conditional probability for each term:
  $P(Forest|c) = (5+1)/(8+6) = 3/7$
- $P(Mad|c) = P(Hatter|c) = (0+1)/(8+6) = 1/14$
- $P(Forest|c_\neg) = (1+1)/(3+6) = 2/9$
- $P(Mad|c_\neg) = P(Hatter|c_\neg) = (1+1)/(3+6) = 2/9$

Our test document then:

- $P(c|d_5) = 3/4 * (3/7)^3 * (1/14)^2 = 0.0003$
- $P(c_\neg|d_5) = 1/4 * (2/9)^3 * (2/9)^2 = 0.0001$
- **the test document belongs to Winnie the Pooh**

NB excercise

Some details:

- The +1 in the nominator is called *Laplace smoothing* and it helps avoid 0 probability (see second row above)
- 8 and 3 are the lenght of the respective texts in each category, and 6 is the unique word count constant.

Support Vector Machine

- Developed in 1990s
- The SVM is the generalization of the maximal margin classifier
- Very flexible, flavour of the decade
- Performs well, "*considered one of the best "out of the box" classifiers*" (ISL, p.337)
- The SVM's goal is to locate the hyperplane with the largest margin (maximum marginal hyperplane)

Goal of SVM

Equation for a separating hyperplane: $W * X + b$ where **W** is a weight vector $W = \{w_1, w_2, ..., w_n\}$ for n attributes and b is a scalar, called bias.

Goal is to find a set of weights that specify two hyperplanes:

$$\vec{w} * \vec{x} + b \geq +1$$

$$\vec{w} * \vec{x} + b \leq +1$$

Maximizing the margin is minimizing $||w||$ or $||w||^2$ (Euclidean distance)

SVM

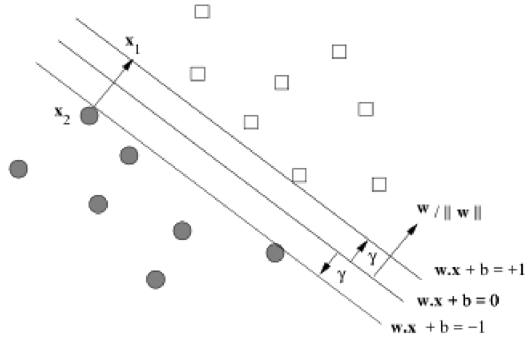


Figure 12.15: Normalizing the weight vector for an SVM

Leskovec, Rajaraman, Ullman (2014), p.463

SVM parameters

- loss function C, which regulates the tolerance to misclassification
- if data is not linearly separable: different kernels (radial, polynomial or sigmoid)

SVM illustrations

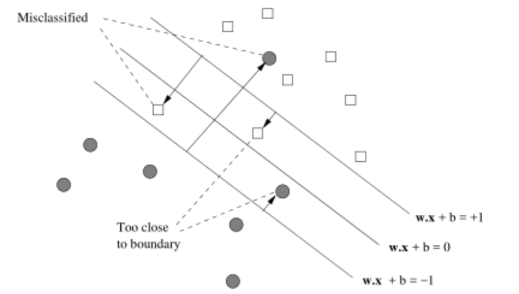
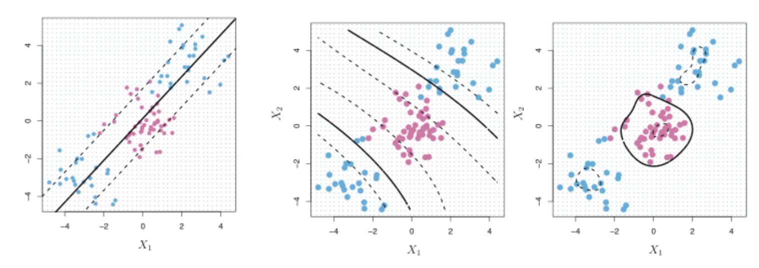


Figure 12.16: Points that are misclassified or are too close to the separating hyperplane incur a penalty; the amount of the penalty is proportional to the length of the arrow leading to that point

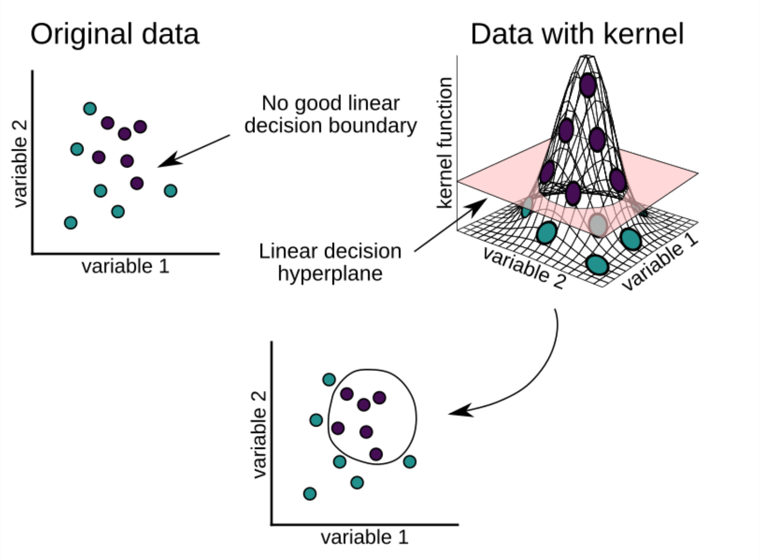
Leskovec, Rajaraman, Ullman (2014), p.465

SVM illustrations



James, Witten, Hastie and Tibshirani (2017), p.353

SVM illustrations



https://machinelearningwithmlr.wordpress.com/2019/10/10/example-post/

Additional materials in this session

Han, Kamber, Pei: Data Mining - Conpcets and Techniques (Third edition), 2012 (Ch8)

Manning, Raghavan, Schütze: Introduction to Information Retrieval, 2008 (Ch 13)

James, Witten, Hastie, Tibshirani: An Introduction to Statistical Learning, 2017 (Ch2, Ch4)

Leskovec, Rajaraman, Ullman: Mining of Massive Datasets, 2014 (Ch12.3)