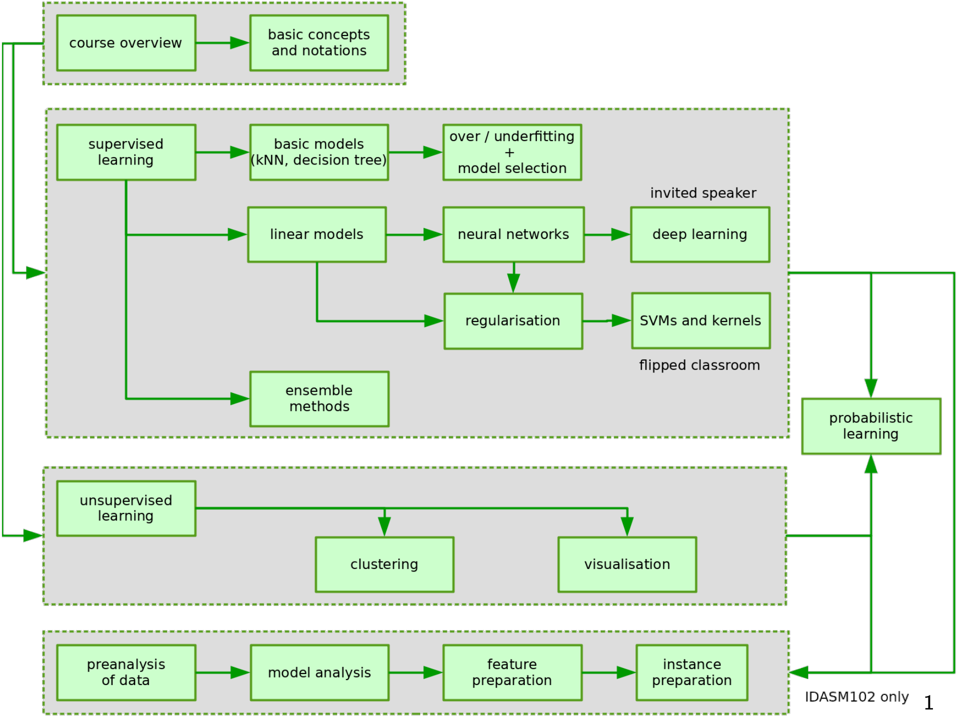
Machine Learning: Lesson 2

Basic Concepts and Notations

Benoît Frénay - Faculty of Computer Science

UNIVERSITÉ DE NAMUR


course overview
basic concepts and notations
supervised learning
basic models (kNN, decision tree)
over / underfitting + model selection
invited speaker
linear models
neural networks
deep learning
regularisation
SVMs and kernels
flipped classroom
ensemble methods
probabilistic learning
unsupervised learning
clustering
visualisation
preanalysis of data
model analysis
feature preparation
instance preparation
IDASM102 only

Outline of this Lesson

- data, models and learning
- example: linear regression
- example: text classification

Data, Models and Learning

Working with Data: Notations

What is an instance?

an instance is an object of interest, generated by some unknown process to be modelled, and often characterised by a set of d features $x_1, x_2 \dots x_d$

- does not fit all possible cases (e.g. proteins, texts, social networks, etc.), but sufficient in many cases (see deep learning and kernels)

What is a dataset?

dataset = set of instances/data to be used for learning

- supervised dataset $\mathcal{D} = \{(\mathbf{x}_i, t_i)\}$
- unsupervised dataset $\mathcal{D} = \{(\mathbf{x}_i)\}$

where $1 \leq i \leq n$ for a dataset of size n ($= n$ instances) and

- $\mathbf{x}_i$ = features of ith object (e.g. picture, medical report, profile, etc.)
- t_i = ith target (e.g. class, only in supervised learning)

Working with Data: Notations

What is an instance?

an instance is an object of interest, generated by some unknown process to be modelled, and often characterised by a set of d features $x_1, x_2 \ldots x_d$

- does not fit all possible cases (e.g. proteins, texts, social networks, etc.), but sufficient in many cases (see deep learning and kernels)

What is a dataset?

dataset = set of instances/data to be used for learning

- supervised dataset $\mathcal{D} = \{(\mathbf{x}_i, t_i)\}$
- unsupervised dataset $\mathcal{D} = \{(\mathbf{x}_i)\}$

where $1 \leq i \leq n$ for a dataset of size n ($= n$ instances) and

- $\mathbf{x}_i$ = features of ith object (e.g. picture, medical report, profile, etc.)
- t_i = ith target (e.g. class, only in supervised learning)

Working with Data: Notations

Vectorial form of dataset

a dataset $\mathcal{D} = \{(\mathbf{x}_i, t_i)\}$ of size n can be written in matrix form with

- a data/design matrix (think of it as an Excel spreadsheet)

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

- a vector of target values (only in supervised learning)

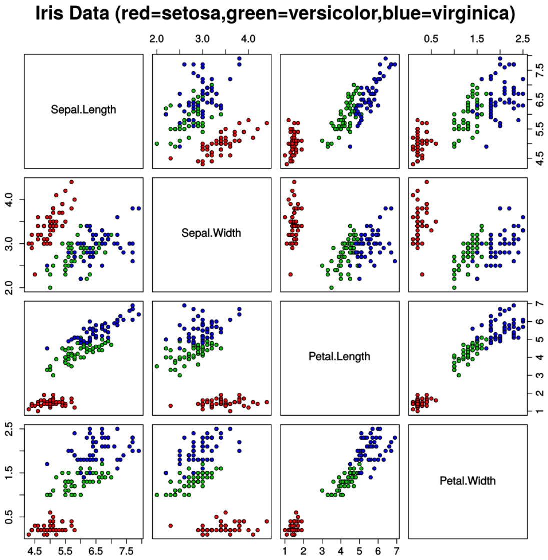
$$\mathbf{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{pmatrix}$$

Example of Simple Dataset: Fisher's Iris flower dataset

Raw data ($n = 150$ flowers of three species)

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	I. setosa
4.9	3.0	1.4	0.2	I. setosa
4.7	3.2	1.3	0.2	I. setosa
...	...	...	...	...
7.0	3.2	4.7	1.4	I. versicolor
6.4	3.2	4.5	1.5	I. versicolor
6.9	3.1	4.9	1.5	I. versicolor
...	...	...	...	...
6.3	3.3	6.0	2.5	I. virginica
5.8	2.7	5.1	1.9	I. virginica
7.1	3.0	5.9	2.1	I. virginica
...	...	...	...	...

Example of Simple Dataset: Fisher's Iris flower dataset



source: ***https://en.wikipedia.org/wiki/Iris_flower_data_set***

From data to Knowledge: Models

Definition

A model provides an approximation to the process which generates data.

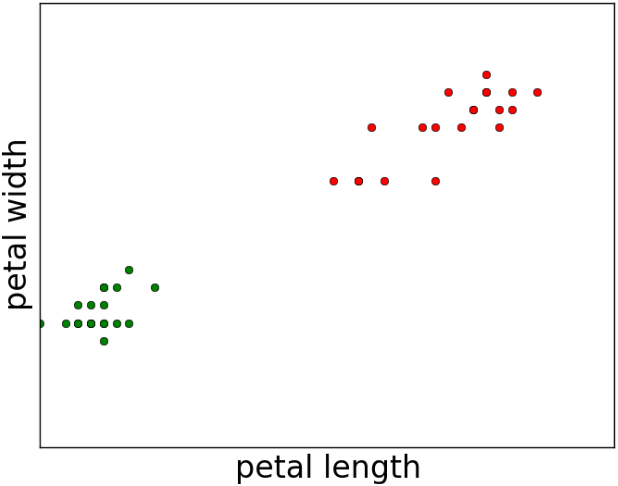
Specificity of machine learning

- we only observe data, i.e. no direct access to the process
- learning consists in building a good and useful approximation
- similar to the identification problem in control and system theory

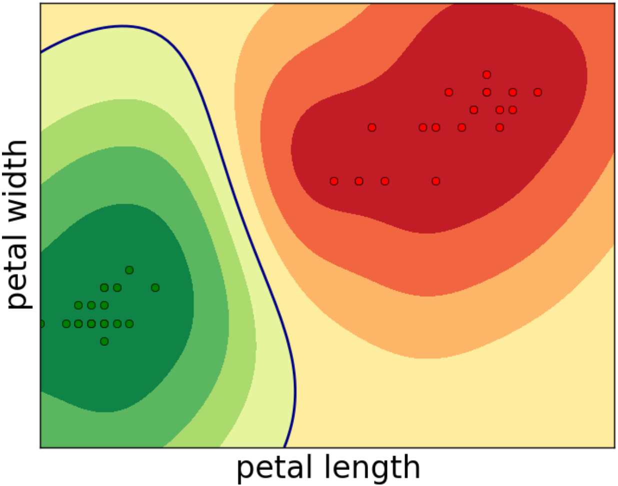
Characteristics of models

- a model is characterised by its parameters (e.g. weights)
- learning = running an algorithm to optimise the parameters

Example of Simple Model for the Iris dataset



Example of Simple Model for the Iris dataset



Types of Models

What is the task ?

- classification: partition objects into several classes (discrete output)
  - examples: healthy/ill patients, spam/non-spam, star/quasar
- regression: associate objects with quantity of interest (real output)
  - examples: diabetes progression, foetus length, photometric redshift
- clustering: find groups (=clusters) of objects (unknown output)
  - examples: groups of clients, stars, texts, products, students
- many other tasks: density estimation, visualisation, recommendation, graph mining, anomaly detection, image segmentation, log analysis...

What is the goal ?

- predictive: make predictions about future (unseen) objects
  - example: does my new patient have breast cancer ?
- descriptive: gain domain knowledge from observed data
  - example: which genes are related to breast cancer ?

Types of Models

What is the task ?

- classification: partition objects into several classes (discrete output)
  - examples: healthy/ill patients, spam/non-spam, star/quasar
- regression: associate objects with quantity of interest (real output)
  - examples: diabetes progression, foetus length, photometric redshift
- clustering: find groups (=clusters) of objects (unknown output)
  - examples: groups of clients, stars, texts, products, students
- many other tasks: density estimation, visualisation, recommendation, graph mining, anomaly detection, image segmentation, log analysis...

What is the goal ?

- predictive: make predictions about future (unseen) objects
  - example: does my new patient have breast cancer ?
- descriptive: gain domain knowledge from observed data
  - example: which genes are related to breast cancer ?

What does it Mean for a Machine to Learn?

Learning means to **find** a model of data

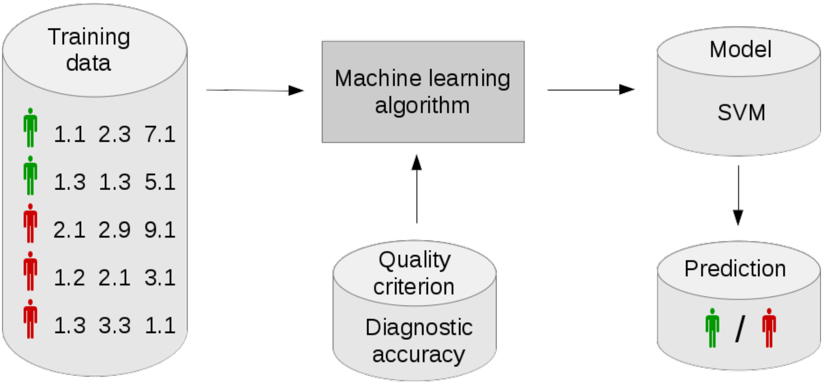
- machine learning studies **how machine can learn automatically**

Three steps

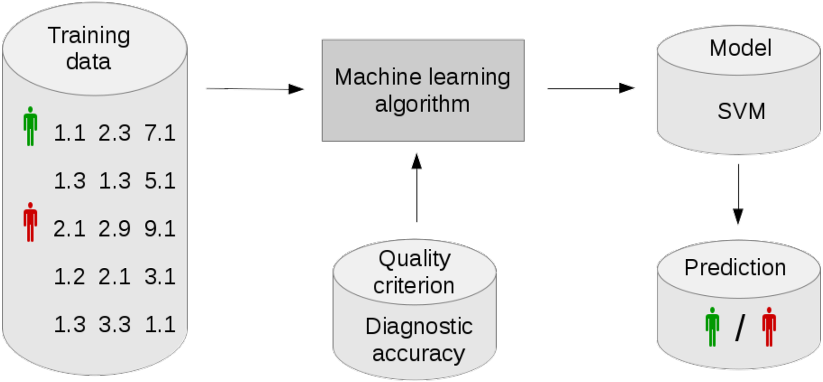
- specify a **type of model** (e.g. a linear model) **what you can get**
- specify a **criterion** (e.g. mean square error) **what you want to get**
- **find** the **best model** w.r.t. the criterion **how you get it**

human are necessary in each step (expert knowledge and algorithm design)

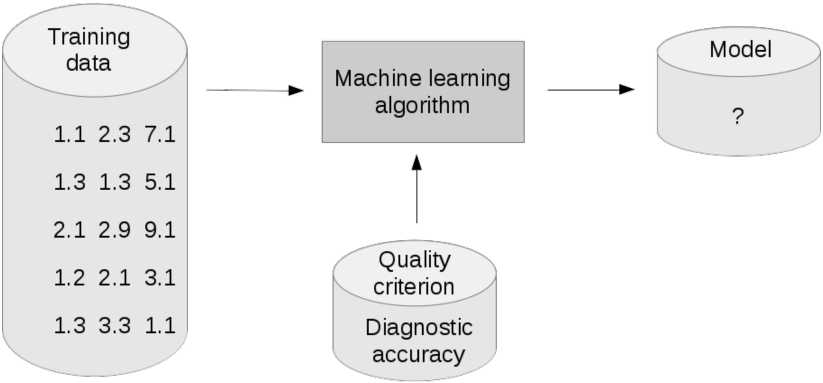
Supervised Machine Learning



Semi-supervised Machine Learning



Unsupervised Machine Learning



First Example: Linear Regression

Example of Learning Process: Linear Regression

Model: linear model (w_j = weight of jth feature + w_0 = bias)

$$f(x_1, \dots, x_d) = w_1 x_1 + \cdots + w_d x_d + w_0$$

prediction = sum of feature values x_j, weighted by w_j (positive or negative)

Criterion: mean square error

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(t_i - f(x_{i1}, \dots, x_{id})\right)^2$$

Algorithm: linear regression / ordinary least squares

Input: dataset $\mathcal{D} = \{(\mathbf{x}_i, t_i)\}$ in matrix/vectorial form as $\mathbf{X}$ and $\mathbf{t}$
Output: optimal weights for linear regression (w.r.t. MSE)

return $\mathbf{w} = \arg\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \left(t_i - f(x_{i1}, \dots, x_{id})\right)^2 = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{t}$

Example of Learning Process: Linear Regression

Model: linear model (w_j = weight of jth feature + w_0 = bias)

$$f(x_1, \ldots, x_d) = w_1 x_1 + \cdots + w_d x_d + w_0$$

prediction = sum of feature values x_j, weighted by w_j (positive or negative)

Criterion: mean square error

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (t_i - f(x_{i1}, \ldots, x_{id}))^2$$

Algorithm: linear regression / ordinary least squares

Input: dataset $\mathcal{D} = \{(\mathbf{x}_i, t_i)\}$ in matrix/vectorial form as $\mathbf{X}$ and $\mathbf{t}$
Output: optimal weights for linear regression (w.r.t. MSE)

return $\mathbf{w} = \arg\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} (t_i - f(x_{i1}, \ldots, x_{id}))^2 = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{t}$

Example of Learning Process: Linear Regression

Model: linear model (w_j = weight of jth feature + w_0 = bias)

$$f(x_1, \ldots, x_d) = w_1 x_1 + \cdots + w_d x_d + w_0$$

prediction = sum of feature values x_j, weighted by w_j (positive or negative)

Criterion: mean square error

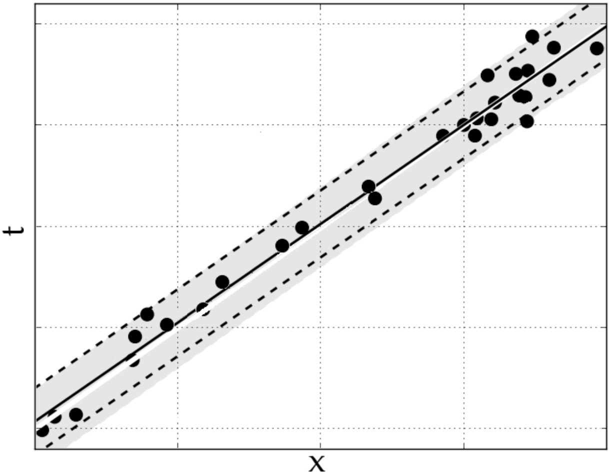
$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (t_i - f(x_{i1}, \ldots, x_{id}))^2$$

Algorithm: linear regression / ordinary least squares

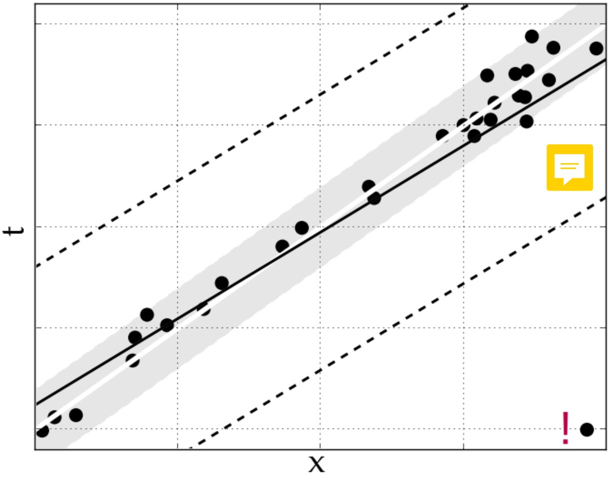
Input: dataset $\mathcal{D} = \{(\mathbf{x}_i, t_i)\}$ in matrix/vectorial form as $\mathbf{X}$ and $\mathbf{t}$
Output: optimal weights for linear regression (w.r.t. MSE)

return $\mathbf{w} = \arg\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} (t_i - f(x_{i1}, \ldots, x_{id}))^2 = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{t}$

Example of Linear Regression



Example of Linear Regression



Application: Diabetes Progression

Task description

- goal: predict the **diabetes progression** one year after baseline
- 442 **diabetes patients** were measured on **10 baseline variables**

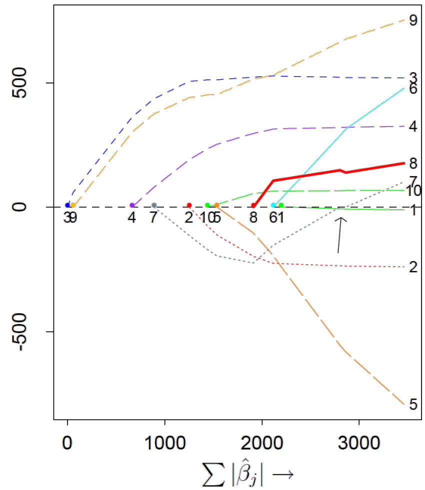
Available patient characteristics (features)

1. age
2. sex
3. body mass index (BMI)
4. blood pressure (BP)
5. serum measurement #1

... ...

10. serum measurement #6

Application: Diabetes Progression

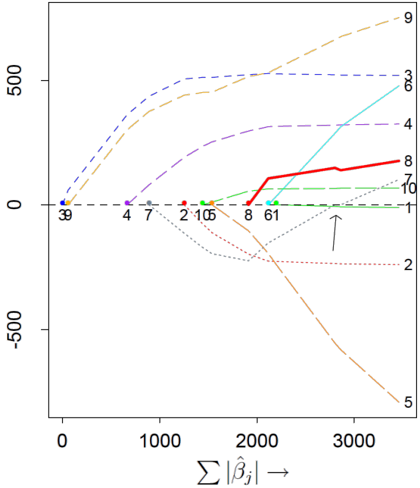


What are the **best features**?

3 body mass index (BMI)
9 serum measurement #5
4 blood pressure (BP)
7 serum measurement #3
2 sex
10 serum measurement #6
5 serum measurement #1
8 serum measurement #4
6 serum measurement #2
1 age

Efron, B., Hastie, T., Johnstone, I., Tishirani, R. Least Angle Regression. Annals of Statistics 32 p. 407–499, 2004.

Application: Diabetes Progression



What are the **1 best features**?

3 body mass index (BMI)

9 serum measurement #5

4 blood pressure (BP)

7 serum measurement #3

2 sex

10 serum measurement #6

5 serum measurement #1

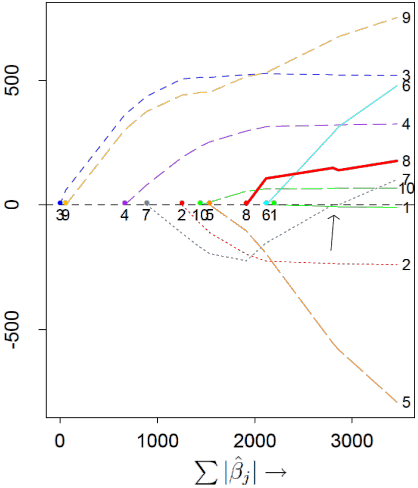
8 serum measurement #4

6 serum measurement #2

1 age

Efron, B., Hastie, T., Johnstone, I., Tishirani, R. Least Angle Regression. Annals of Statistics 32 p. 407–499, 2004.

Application: Diabetes Progression



What are the **2 best features**?

3 body mass index (BMI)

9 serum measurement #5

4 blood pressure (BP)

7 serum measurement #3

2 sex

10 serum measurement #6

5 serum measurement #1

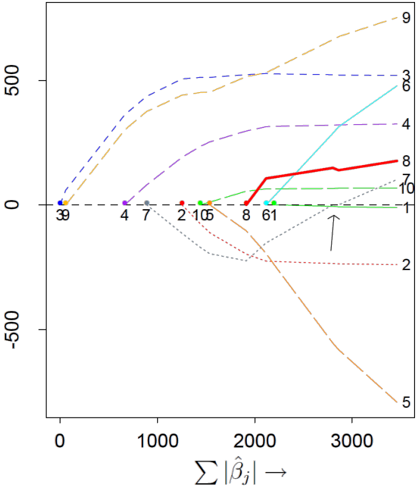
8 serum measurement #4

6 serum measurement #2

1 age

Efron, B., Hastie, T., Johnstone, I., Tishirani, R. Least Angle Regression. Annals of Statistics 32 p. 407–499, 2004.

Application: Diabetes Progression



What are the **3 best features**?

3 body mass index (BMI)

9 serum measurement #5

4 blood pressure (BP)

7 serum measurement #3

2 sex

10 serum measurement #6

5 serum measurement #1

8 serum measurement #4

6 serum measurement #2

1 age

Efron, B., Hastie, T., Johnstone, I., Tishirani, R. Least Angle Regression. Annals of Statistics 32 p. 407–499, 2004.

Second Example: Text Classification

Text Preprocessing in a Nutshell

Issues related to text classification

- naive approach : count words in documents to get features
- most common words are stop words (*the*, *and*, etc.)
- rare words are not necessarily relevant terms

TF-IDF: how important term i is to a document j in the corpus

$$TF.IDF_{ij} = TF_{ij} \times IDF_i$$

where

- $TF_{ij} = f_{ij} / \max_k f_{kj}$ = term frequency (TF) of term i in document j
- $IDF_i = \log_2(n/n_i)$ = inverse document frequency (IDF) in corpus

Text Preprocessing in a Nutshell

Issues related to text classification

- naive approach : count words in documents to get features
- most common words are stop words (*the*, *and*, etc.)
- rare words are not necessarily relevant terms

TF-IDF: how important term i is to a document j in the corpus

$$TF.IDF_{ij} = TF_{ij} \times IDF_i$$

where

- $TF_{ij} = f_{ij} / \max_k f_{kj}$ = term frequency (TF) of term i in document j
- $IDF_i = \log_2 (n/n_i)$ = inverse document frequency (IDF) in corpus

Application: Reuters Stories Classification

The Reuters Corpus Volume I (RCV1) dataset

- an archive of over 800,000 manually categorised Newswire stories
- made available by Reuters, Ltd. for research purposes
- topic codes capture the major subject of each story
  - CCAT (corporate/industrial)
  - ECAT (economics)
  - GCAT (government/social)
  - MCAT (markets)

Classification task

- classify documents in the CCAT category (or not)
- training set and test set contain 781,265 and 23,149 instances
- 47,152 TF/IDF features were computed for this task
- learning with a simple linear model (hinge/logistic loss SVM)

Application: Reuters Stories Classification

Example of big data problem

- 781,265 instances $\times$ 47,152 features $=$ 36,838,207,280 values
- you **need** to use sparse data structures, fast algorithms, etc.

Results obtained by state-of-the-art solvers

support vector machines with hinge/logistic loss

algorithm	training time	test error
SVMLight	23,642 s	6.02%
SVMPerf	66 s	6.03%
SGD	1.4 s	6.02%

algorithm	training time	test error
TRON	30 s	5.68%
SGD	2.3 s	5.66%

Outline of this Lesson

- data, models and learning
- example: linear regression
- example: text classification