Selected Topics in Applied Machine Learning: An integrating view on data analysis and learning algorithms

ESSLLI '2015
Barcelona, Spain

http://ufal.mff.cuni.cz/esslli2015

Barbora Hladká
hladka@ufal.mff.cuni.cz

Martin Holub
holub@ufal.mff.cuni.cz

Charles University in Prague,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

Welcome to the lessons!

Course web page: `http://ufal.mff.cuni.cz/esslli2015`

- All materials will be available at the web page
  - additional links
  - list of references
  - . . . etc.
- We will post every day after the lesson
  - presented slides
  - data for your experiments and demo R scripts
  - materials needed for your homeworks
- Course is organized in blocks. Please, ask questions between blocks.

Block 1.1
Introduction to the lessons

Purpose of this intro – why we need it?

- **Overview of the expected knowledge and prerequisites**
  - elementary concepts of machine learning
  - necessary maths
  - fundamental knowledge of R and practical NLP

- **Our terminology and an example machine learning task**

Focus of the course – brief outline

- **Data analysis**
  – deeper understanding ML task by statistical view on data
- **Ensemble learning methods**
  – combining multiple learners, sampling, bagging, boosting
  – AdaBoost, Random Forests
- **Model complexity and regularization**
  – underfitting, overfitting, and regularization
- **Feature selection**
  – measures of feature relevance and feature selection algorithms
- **Model assessment and selection**
  – evaluating a model's performance
  – selecting the proper level of flexibility for a model

Main goals of the course

This introductory course is aimed mainly at the people who need/want to practically apply machine learning methods in the NLP area. It should help beginners to deepen their understanding.

Our main goals

- Deeper understanding of the ML process
- Experience with complex ML experiments
- Practical steps towards model selection

Two recommendations for real beginners

1) If you are not really certain about your fundamental knowledge of ML, our **"fundamental" course at ESSLLI '2013** would be helpful.
— see `http://ufal.mff.cuni.cz/mlnlpr13`

- That course was intended for real beginners. This time we suppose that you students are familiar with the subject as it was presented in 2013, and our goal is to broaden your horizons.

2) Also, we can recommend our recent article written mainly for students *"A Gentle Introduction to Machine Learning for Natural Language Processing – How to start in 16 practical steps"*, which covers the ML fundamentals with focus on using ML in the NLP area.

- In this course we will refer to this paper as **"16 steps"**. For your personal perusal, you can obtain a copy of this paper from us upon your specific request.

Example task: “Movie recommendations” (MOV)

There is a publicly available data base called “*Movie lens*” containing 1) data about movies, 2) data about users, and 3) users’ ratings.

Typically, users give their votes only for a small number of movies that they know.

Excerpt from the data – about users and movies

	age	**gender**	**occupation**	**zip code**
Peter	19	M	student	58644
Mary	50	F	healthcare	60657

title	**action**	...	**IMDb rating**	**director**
Toy Story	0	...	8.3	John Lasseter
Some Like It Hot	0	...	8.3	Billy Wilder
Star Wars	1	...	8.7	George Lucas

MOV – users' ratings to be predicted

Excerpt from the data – users' ratings

	Toy Story (1995)	Star Wars (1977)	Some Like It Hot (1959)
Peter	**?**	5	4
Paul	2	5	**?**
Mary	2	4	**?**

User's rating is the value that should be automatically predicted, for a given user and a given movie.

– E.g., predict Mary's rating for the movie *Some Like it Hot*.

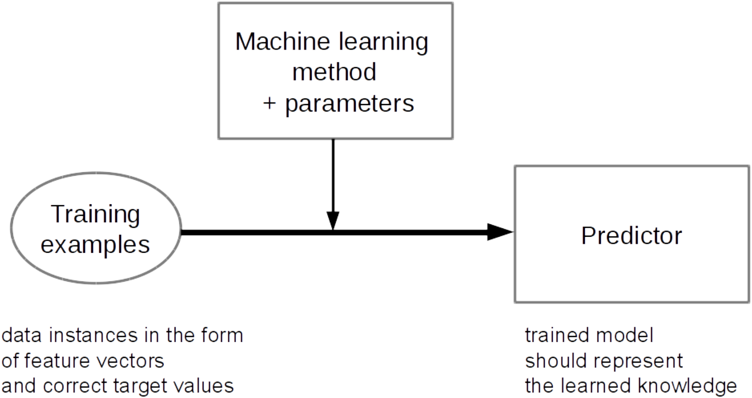
MOV – summary of the task characteristics

Any supervised machine learning task
is characterized by the data available for learning

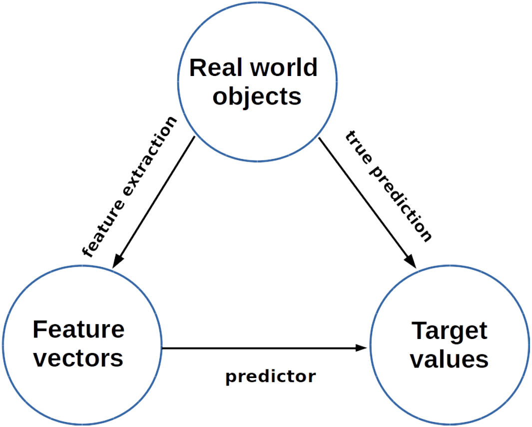
- **MOV examples** – Set of records, each consisting of a user, a movie, and the user's rating
- **MOV task** – Predict the user's rating for a given movie.
  E.g., predict Mary's rating for *Star Wars*.
- **Target values** – users' ratings between 1 to 5
- **Feature vectors** consist of
  - data about users (5 features)
  - data about movies (26 features)

Supervised learning process

Supervised Machine Learning = computer learns "essential knowledge" extracted from a (large) set of examples



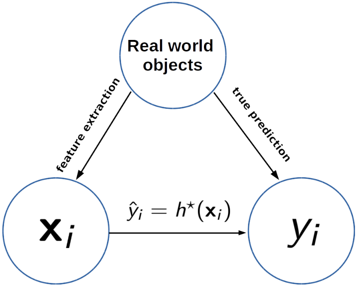
Machine learning as building a prediction function



- if target values are *continuous* numbers, we speak about **regression**
  = estimating or predicting a continuous response
- if target values are *discrete/categorical*, we speak about **classification**
  = identifying group membership

Prediction function and its relation to the data

Idealized model of supervised learning



- $\mathbf{x}_i$ are **feature vectors**, y_i are true **predictions**
- **prediction function** $h^\star$ is the "best" of all possible hypotheses h
- **learning process** is searching for $h^\star$, which means to search the **hypothesis space** and minimize a predefined **loss function**
- ideally, the learning process results in $h^\star$ so that predicted $\hat{y}_i = h^\star(\mathbf{x}_i)$ is equal to the true target values y_i

Loss function

A loss function $\mathrm{L}(\hat{y}, y)$ measures the cost of predicting $\hat{y}$ when the true value is y. Commonly used loss functions are

- squared loss $\mathrm{L}(\hat{y}, y) = (\hat{y} - y)^2$
  for regression
- zero-one loss $\mathrm{L}(\hat{y}, y) = \mathrm{I}(\hat{y}_i \neq y_i)$
  for classifiation; *indicator variable* I is 1 if $\hat{y}_i \neq y_i$, 0 otherwise

The goal of learning can be stated as producing a model with the smallest possible loss; i.e., a model that minimizes the average $\mathrm{L}(\hat{y}, y)$ **over all examples.**

Note: loss function is sometimes also known as "cost function".

Supervised ML task and data instances

Supervised machine learning necessarily requires learning examples

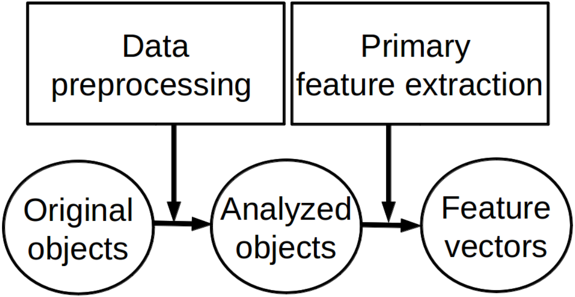
- **Features** are properties of examples that can be observed or measured
  – are numerical (discrete or continuous), or categorical (incl. binary)
- **Feature vector** is an ordered list of selected features
- **Data instance** = feature vector (+ target class, if it is known)
- **Training data** = a set of examples used for **learning process**
- **Test data** = another set of examples used for **evaluation**

Terminology – features and target values

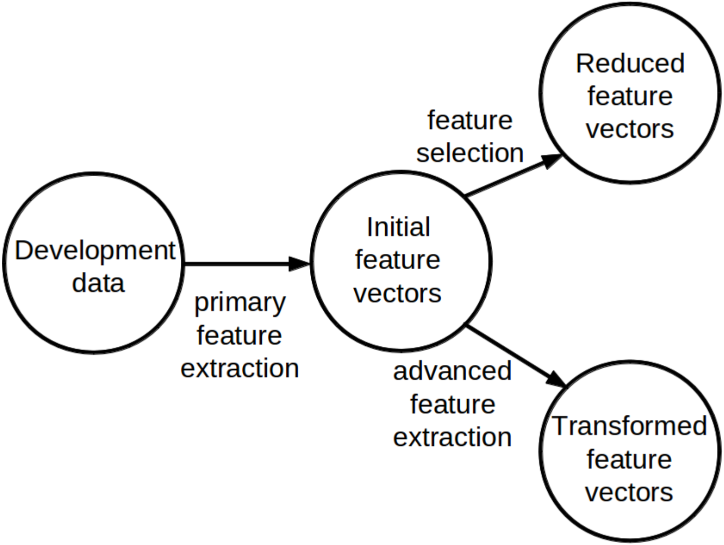
- **How different people call values that describe objects**

	observed (known) object characteristics	**values or categories to be predicted**
computer scientists	**features**	**(target) value or class**
mathematicians (statisticians)	attributes or predictors	response (value) or output value

Data preprocessing and feature extraction



Feature extraction and feature selection



Sample error and generalization error

Sample error of a hypothesis h with respect to a data sample S of the size n is usually measured as follows

- for **regression**: **mean squared error** $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$
- for **classification**: **classification error** $= \frac{1}{n}\sum_{i=1}^{n}\mathrm{I}(\hat{y}_i \neq y_i)$

Generalization error (aka "true error" or "expected error") measures how well a hypothesis h generalizes beyond the used training data set, to unseen data with distribution $\mathcal{D}$. Usually it is defined as follows

- for **regression**: $\mathrm{error}_{\mathcal{D}}(h) = \mathsf{E}\,(\hat{y}_i - y_i)^2$
- for **classification**: $\mathrm{error}_{\mathcal{D}}(h) = \Pr\,(\hat{y}_i \neq y_i)$

Accuracy and error rate

To measure the performance of classification tasks we often use (sample) *accuracy* and (sample) *error rate*

Sample accuracy is the number of correctly predicted examples divided by the number of all examples in the predicted set

Sample error rate is equal to **1 - accuracy**

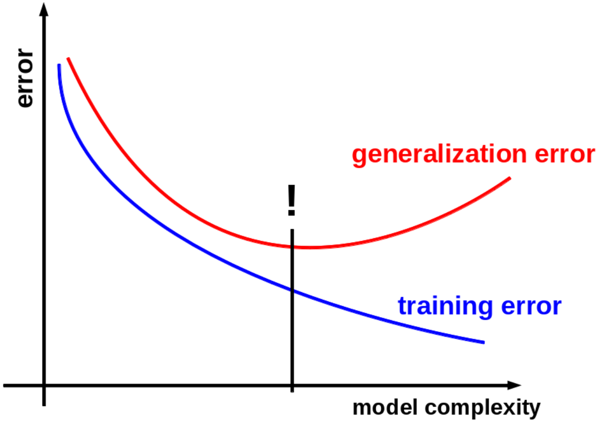
Training error rate is the sample error rate measured on the training data set

Test error rate is the sample error rate measured on the test data set

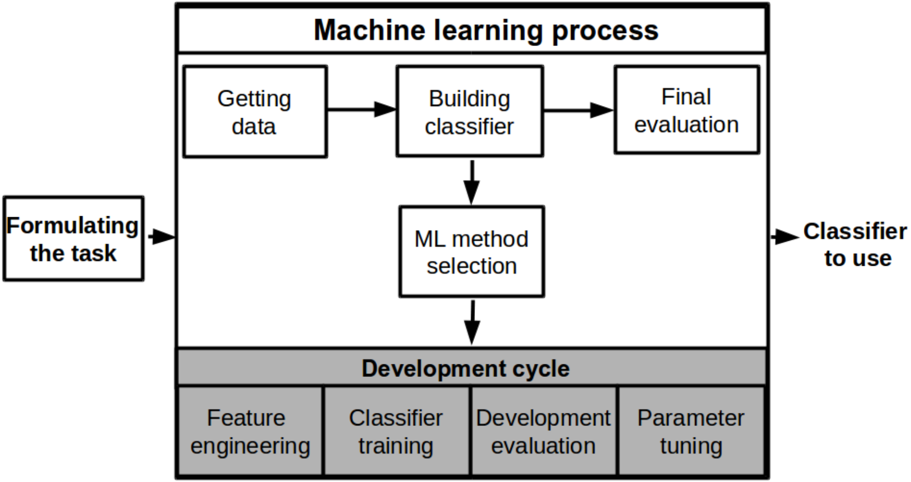
Minimizing generalization error

Finding a model that minimizes generalization error

. . . is one of central goals of the machine learning process



Machine learning process – development cycle



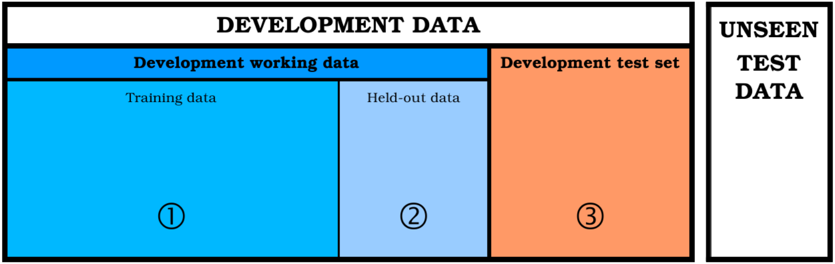
Terminological note on building predictors

The purpose of the learning process is search for the best prediction function parameters

learning parameters	hypothesis parameters
= parameters of the learning process	= parameters of the prediction function

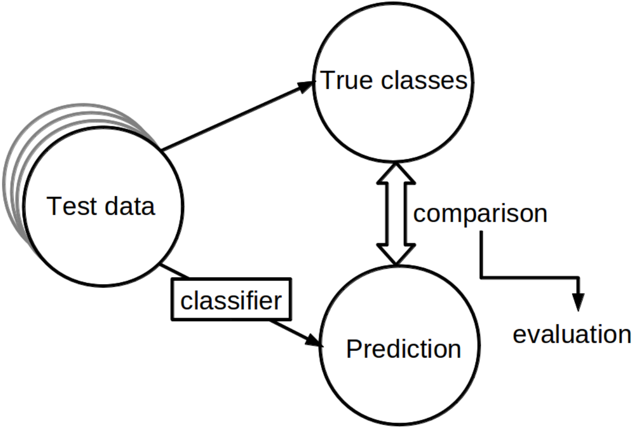
- **Method** = approach/principle to learning. i.e. to building predictors
- **Model** = method + set of features + learning parameters
- **Predictor** = trained model, i.e. an output of the machine learning process, i.e. a particular method trained on a particular training data.
- **Prediction function** = predictor (used in mathematics). It's a function calculating a response value using "predictor variables".
- **Hypothesis** = prediction function – not necessarily the best one (used in theory of machine learning).

Development data and its division



All subsets should be selected randomly in order to represent the characteristic distribution of both feature values and target values in the available set of examples.

Evaluation – basic scheme



Evaluation – overview of elementary concepts

You should know

- k-fold cross-validation, leave-one-out cross-validation
- stratified cross-validation
  - if subsets are built so as to preserve the original class distribution in all subsets

Related to classification:

- confusion matrices
- accuracy, precision, recall, F-measure

Examples of learning methods

- Decision Trees (DT)
- Naïve Bayes classifier (NB)
- Support Vector Machines (SVM)
- Logistic Regression (LogR)
- k Nearest Neighbours (kNN)

Probability and statistics – necessary knowledge

- difference between **populations and samples**
    - population parameters vs. sample statistics
- discrete and continuous **random variables**
- **probability mass function** (PMF) – also called density function
- **expected value, variance, standard deviation**
- how we describe **probabilistic distributions**
    - cumulative distribution function (CDF)
    - probability density function (PDF)
    - quantile function (QF)
- **normal distibution, binomial distribution**
- **conditional probability**
- **statistical independence**

Information theory – entropy

The average amount of information that you get when you observe discrete/categorical random values is

$$\boxed{-\sum_{value} \Pr(value) \cdot \log_2 \Pr(value)}$$

This is what information theory calls entropy.

- Entropy of a random variable X is denoted by $\mathrm{H}(X)$.
- Entropy is a measure of the uncertainty in a random variable.

The unit of entropy is bit. Entropy says how many bits on average you necessarily need to encode a value of the given random variable.

Information theory – conditional entropy

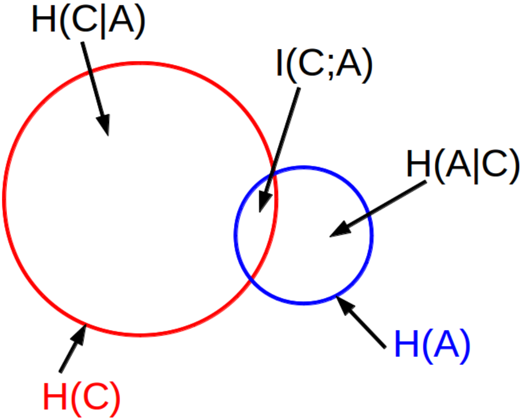
Both target class and discrete features can be measured by entropy. However, their entropy itself does not tell us about their relationship.

How much does target class entropy decrease if we have the knowledge of a discrete feature?

The answer is **conditional entropy**:

$$\boxed{\mathrm{H}(C \mid A) \quad = \quad - \sum_{y \in C,\ x \in A} \Pr(y, x) \cdot \log_2 \Pr(y \mid x)}$$

Conditional entropy and mutual information



WARNING

There are NO SETS in this picture! Entropy is a quantity, only a number!

Conditional entropy and mutual information

Mutual information measures the amount of information about one random variable that can be obtained by observing another.

Mutual information is a symmetrical quantity.

$$\boxed{\mathrm{H}(C) - \mathrm{H}(C \mid A) \quad = \quad \mathrm{I}(C; A) \quad = \quad \mathrm{H}(A) - \mathrm{H}(A \mid C)}$$

Another name for mutual information is **information gain**.

Organizational remarks
Exercises and Homeworks

All exercises mentioned during our course are recommended

Some exercises are called **“Homeworks”**
– which means that we *strongly* recommend to do it

Homework solutions will be posted at the course web page next day

Exercise for today

Check if you are familiar with basic concepts and methods introduced in "16 steps"

— For more details you can also read our slides from ESSLLI '2013
`http://ufal.mff.cuni.cz/mlnlpr13`

Short questions?

- **Movie recommendation task (MOV)**
  Predict the user's rating for a given movie

	Toy Story (1995)	Star Wars (1977)	Some Like It Hot (1959)
Peter	**?**	5	4
Paul	2	5	**?**
Mary	2	4	**?**

E.g., predict Mary's rating for the movie *Some Like it Hot*

MOV – Available data

- **About users**

	age	**gender**	**occupation**	**zip code**
Peter	19	M	student	58644
Mary	50	F	healthcare	60657

- **About movies**

title	**action**	...	**IMDb rating**	**director**
Toy Story	0	...	8.3	John Lasseter
Some Like It Hot	0	...	8.3	Billy Wilder
Star Wars	1	...	8.7	George Lucas

MOV – Getting examples

- Create a database of movies to be rated by users
- Set up a rating scale allowing users to rate movies
- Record users' ratings
- Typically, the dataset of ratings is sparse.
  So do some pruning, like require a minimum of twenty ratings per user

MOV – Data

Basic statistics

number of ratings	100,000
number of movies	1,682
number of users	943

- Data comes from the MovieLens datasets
  – for more details, go to the course web page

MOV – Data representation

	1	2	3	4	5-8	9-33
vote id	MOVIE	USER	RATING	TIMESTAMP	user features	movie features
1	1	1	5	1997-09-23 00:02:38	24 M technician 85711	Toy Story (1995) ...
...	...	...	...	...	...	...
100,000	1682	916	3	...	...	...

- For more details, see **mov.pdf** posted at the course webpage

MOV – Loading the data into R

```
# get examples, votes, movies, users
> source("load-mov-data.R")
> nrow(examples)
[1] 100000

> names(examples)
 [1] "movie"                "user"                 "rating"
 [4] "timestamp"            "age"                  "gender"
...
[31] "directors"            "writers"              "stars"
```

Data analysis

Machine learning process

1. Formulating the task (e.g. predict user's rating for a given movie)
2. Getting data (e.g. MOV data)
    - **Data analysis**
3. Building predictor
4. Evaluation

Data analysis

Deeper understanding the task by statistical view on the data

We exploit the data in order to make prediction of the target value.

- Build intuition and understanding for both the task and the data
- Ask questions and search for answers in the data
  - **What values do we see**
  - **What associations do we see**
- Do plotting and summarizing

Data analysis

We focus on

- Recap of methods for basic data exploration
- Analyzing distributions of values
- Analyzing association between features
- Analyzing association between features and target value

Methods for basic data exploration

Frequency tables display the frequency of categorical feature values.

```
# frequency of men and women voting
> table(examples$gender)
  F   M
25740 74260
```

Methods for basic data exploration

Bar plots visualize frequency tables

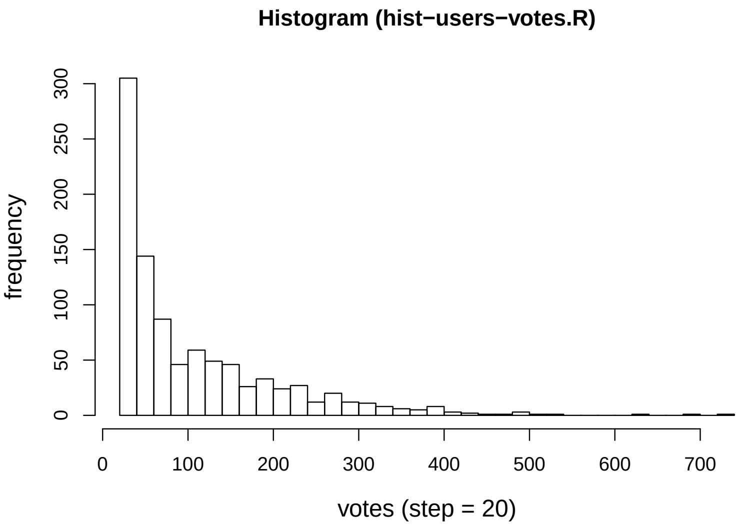
Methods for basic data exploration

Histograms visualize distribution of feature values.

Add a new feature VOTES for the number of votes of the users

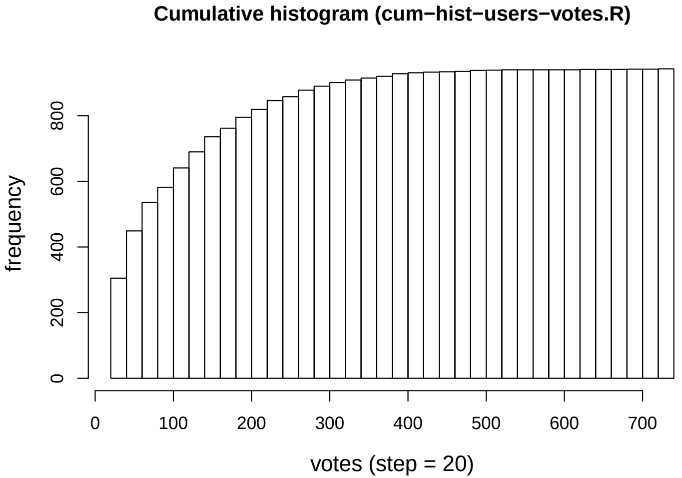
```
# get the number of votes for each user
> user.id <- users[,1]
> votes <- rowSums(sapply(votes[,1],
                 function(x) x==user.id))
> users$votes <- votes
> min(users$votes)
[1] 20
> max(users$votes)
[1] 737
```

Methods for basic data exploration



Methods for basic data exploration

Cumulative histograms visualize cumulative frequencies.



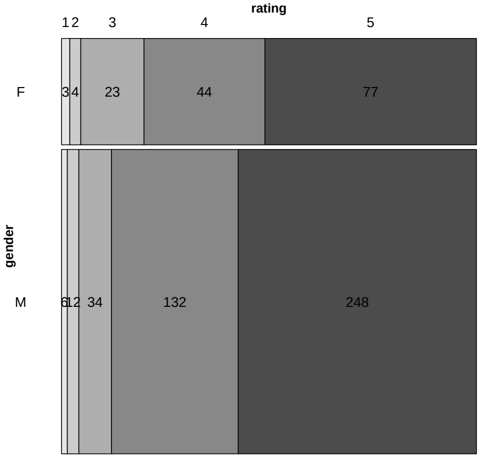
Methods for basic data exploration

Contingency tables display the frequency of values for combination of two categorical features.

```
> # Star Wars ratings
> movie <- subset(examples, movie == 50)
> # construct contingency table
> ct <- table(movie$gender, movie$rating)
> margin.table(ct)                 # total sum
[1] 583
> addmargins(ct)                   # add marginal sums
        1   2   3   4   5 Sum
  F     3   4  23  44  77 151
  M     6  12  34 132 248 432
  Sum   9  16  57 176 325 583
> round(prop.table(ct),2)          # generate proportions
       1    2    3    4    5
  F 0.01 0.01 0.04 0.08 0.13
  M 0.01 0.02 0.06 0.23 0.43
```

Methods for basic data exploration

Mosaic plots visualize contingency tables.

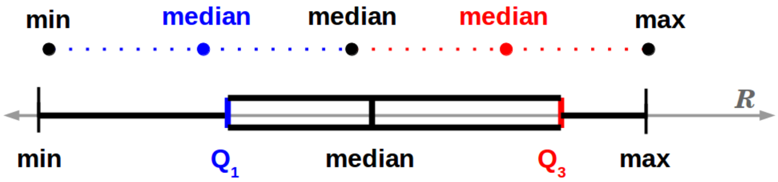


Methods of basic data exploration

Measures of center and variation

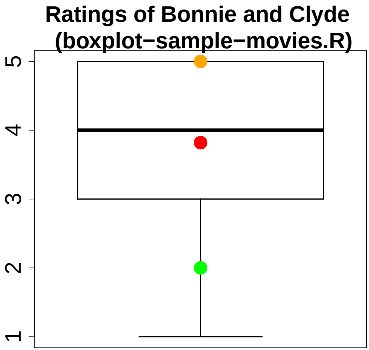
```
> min(users$vote);max(users$vote)
[1] 20
[1] 737
> mean(users$vote)
[1] 106.4
> median(users$vote)
[1] 65
> summary(users$vote) # five-number summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     20      33      65     106     148     737

> sd(users$vote) # standard deviation
[1] 100,93
```

Box-and-whiskers plots visualize five-number summaries.



Methods of basic data exploration

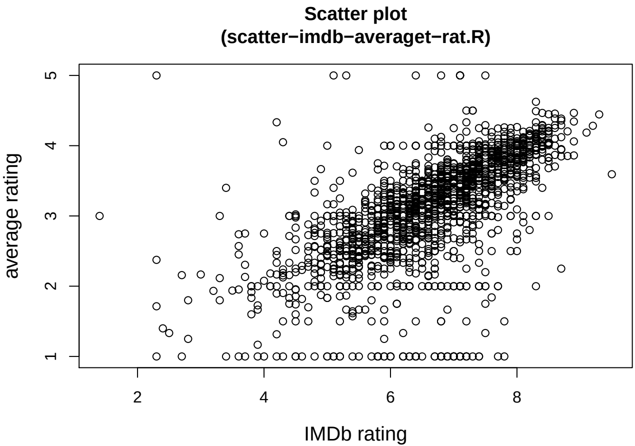
Box-and-whiskers plots



- the average rating is in red, Peter's rating in green and Mary's rating in orange

Methods of basic data exploration

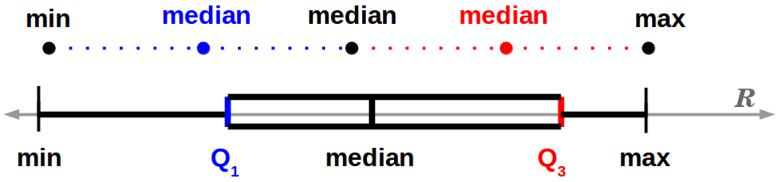
Scatter plots display values of two numerical features.



What values do we see
Analyzing distributions of values

Boxplots are of a great importance to detect outliers

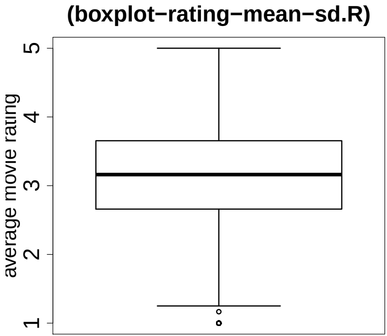
Outlier is an observation that is distant from other observations, typically if it falls more than $1.5 * (Q_3 - Q_1)$ above Q_3 or below Q_1

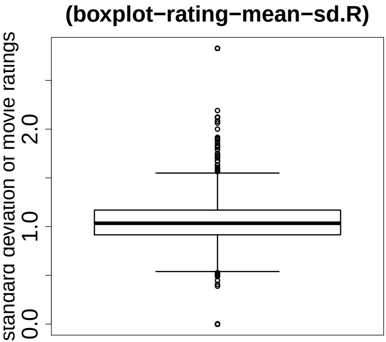


Analyzing distributions of values

Boxplots are of a great importance to detect outliers

We expect that when users are choosing a movie to watch, they check its average rating first and then variance of its ratings.





Analyzing distributions of values

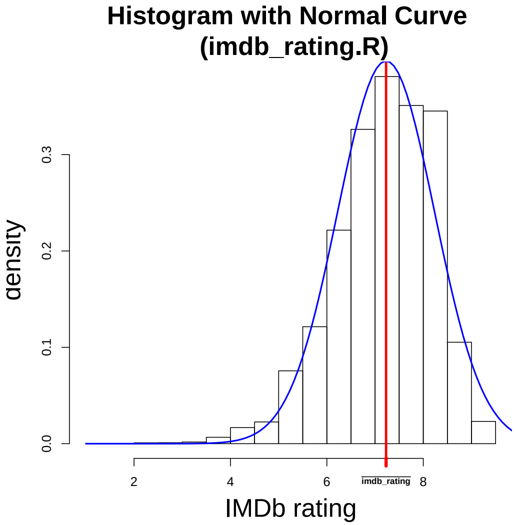
Boxplots are of a great importance to detect outliers

```
> boxplot <- boxplot(tapply(votes$rating, votes$movie, sd))
# analyze outliers
> boxplot$out[1:2]
     247      314
1.788854 0.000000
>
> subset(votes, movie == 247) # Turbo: A Power Rangers Movie (1997)
      user movie rating           timestamp
38147   38   247      5 1998-04-13 03:04:20
38148    1   247      1 1997-09-26 04:40:19
38149  374   247      1 1997-12-01 01:35:22
38150  222   247      1 1997-11-05 08:29:58
38151  782   247      1 1998-04-02 08:48:20
```

Analyzing distributions of values

Analyzing imdb_rating

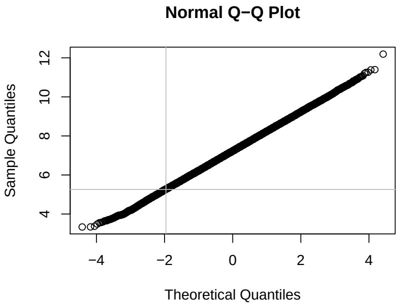
- **What kind of probability distribution characterizes the IMDb ratings?**

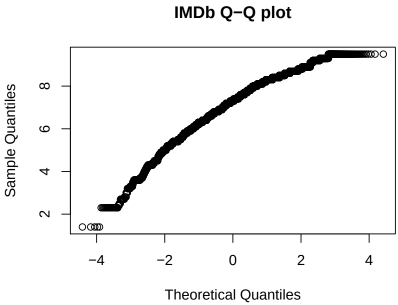


Does IMDB_RATING **follow a normal distribution?**

- Visualize the distribution using a quantile-quantile plot
- Use a distribution test

Visualize the distribution using a quantile-quantile plot





- **Draw a conclusion**: IMDB_RATING does not follow a normal distribution.

Use a distribution test

1. State
   - H_0: IMDB_RATING follows a normal distribution.
   - H_A: IMDB_RATING does not follow a normal distribution.

Use a distribution test

2. Do the **Kolmogorov-Smirnov one-sample test**

```
> ks.test(imdb, "pnorm", mu, sigma)
One-sample Kolmogorov-Smirnov test

data:  imdb
D = 0.0645, p-value < 2.2e-16
alternative hypothesis: two-sided

Warning message:
In ks.test(imdb, "pnorm", mu, sigma) :
ties should not be present
    for the Kolmogorov-Smirnov test
```

Use a distribution test

Ties are observations with the same values.

```
> unique(sort(imdb))
[1] 1.4 2.3 2.4 2.5 2.7 2.8 3.0 3.2 3.3 3.4 3.5 3.6 ...
[20] 4.4 4.5 4.6 4.7 4.8 4.9 5.0 5.1 5.2 5.3 5.4 5.5 ...
[39] 6.3 6.4 6.5 6.6 6.7 6.8 6.9 7.0 7.1 7.2 7.3 7.4 ...
[58] 8.2 8.3 8.4 8.5 8.6 8.7 8.8 8.9 9.1 9.2 9.3 9.5
> length(unique(sort(imdb)))
[1] 69
```

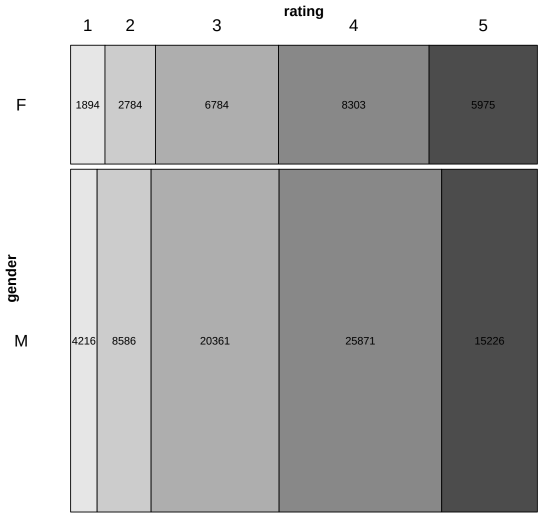
Use a distribution test

```
> ?jitter()
Description:
     Add a small amount of noise to a numeric vector.

> ks.test(jitter(imdb),"pnorm", mu, sigma)
data:  jitter(imdb)
D = 0.0565, p-value < 2.2e-16
alternative hypothesis: two-sided
```

3. Set a significance level $\alpha = 0.05$
4. **Draw a conclusion**: As the *p*-value $< \alpha = 0.05$, we do reject the null hypothesis that IMDB_RATING follows a normal distribution.

Association between feature and target value
Categorical features

Association between gender and rating



Association between feature and target value

Gender and rating

We can see some structure in the mosaic plot. Conduct a statistical test.

1. State
   - H_0: GENDER and RATING are statistically independent.
   - H_A: GENDER and RATING are statistically dependent.
2. Use Pearson's χ^2 test (chi-square test)

```
> ct <- table(examples$gender, examples$rating)
> chisq.test(ct)
Pearson's Chi-squared test
data:  ct
X-squared = 209.1421, df = 4, p-value < 2.2e-16
```

3. Set a significance level $\alpha = 0.05$
4. **Draw a conclusion:** $p < \alpha$ thus we do reject the null hypothesis that RATING and GENDER are statistically independent.

Association between feature and target value
Gender and rating

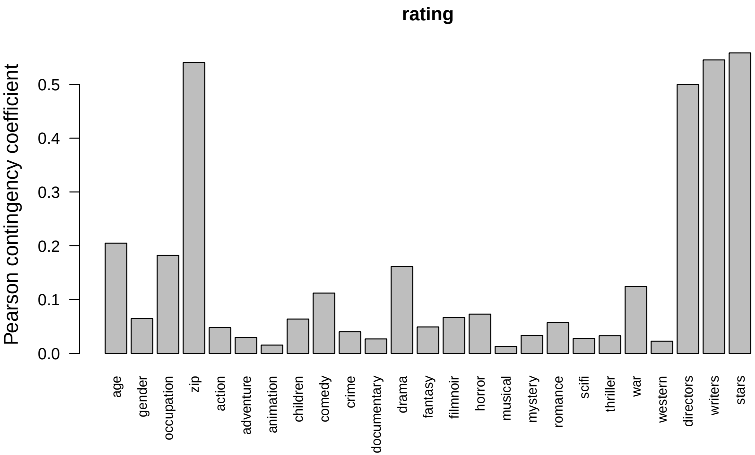
There is some association between GENDER and RATING, i.e. the values of GENDER generally co-occur with certain values of RATING.

What is its strength?

Compute e.g. **Pearson contingency coefficient** (pcc)

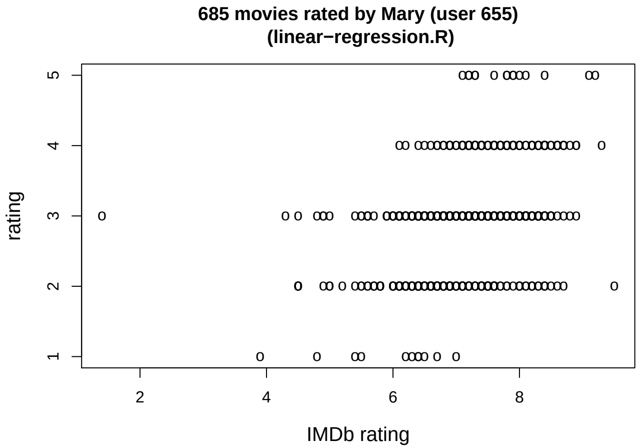
- $0 < \mathrm{pcc} < 1$
- perfect correlation if $\mathrm{pcc} \to 1$
- no correlation if $\mathrm{pcc} \to 0$

Association between feature and target value



Association between feature and target value
Numerical features

Association between Mary's ratings and the IMDb ratings



Association between feature and target value
Numerical features

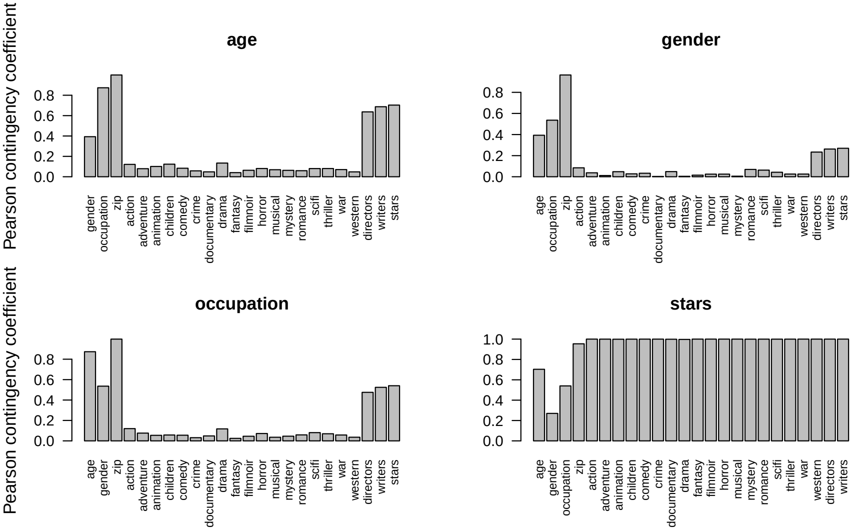
Association between Mary's ratings and the IMDb ratings

Compute e.g. **Pearson correlation coefficient** that is a measure of the linear relationship between features (ρ for a population and r for a sample)

- $-1 \leq r \leq +1$
- perfect negative correlation if $r = -1$
- perfect positive correlation if $r = +1$
- not linear relationship if $r = 0$

$r(\mathrm{Peter's\ rating, imdb\_rating})$	0,51
$r(\mathrm{Paul's\ rating, imdb\_rating})$	0,44
$r(\mathrm{Mary's\ rating, imdb\_rating})$	0,37

Associations between features
Categorical features



Associations between users' rating

Compute e.g. **Pearson correlation coefficient**

$r(\text{Peter's rating}, \text{Mary's rating})$	0,293
$r(\text{Peter's rating}, \text{Paul's rating})$	0,285
$r(\text{Paul's rating}, \text{Mary's rating})$	0,239

Homework 1.2

Work with the MOV data

1. - Get the movies rated at least 3 times
   - Sort them according to their average rating in descending order
   - Focus on the Top 5
     - Which of them has the least variance?
     - Which of them has the highest variance?
     - Which one would you like to see?
2. - Add a new user feature for his/her average rating
   - Does this feature follow a normal distrubution?
   - Visualize the distribution using a Q-Q plot
   - Do the Kolmogorov-Smirnov one sample test
3. - Compute association between genres using Pearson contingency coefficient

Questions?