Basics of Statistics

Lecture 1b

Probability

How likely it is that some event will happen?

Idea:

- Experiment
- Outcomes (sample points) O_1, O_2,… O_n
- Sample space Ω
- Event A
- Probability function P: Events $\rightarrow$[0,1]

Probability

Example: Tossing a coin two times

http://cdn.toonvectors.com/images/35/10267/toonvectors-10267-940.jpg

Example:

- $p(A)$ frequency of observing A
- $p(A, B)$ frequency of observing A and B
- $p(B|A)$ frequency of observing B given A

Properties and definitions

- One can think of events as sets
  - Set operations are defined: $\mathrm{A} \cup B, A \cap B, \bar{A} \backslash B$
- $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$

- Independence $P(A, B) \equiv P(A \cap B) = P(A)P(B)$

- Conditional probability $P(A|B) = \frac{P(A,B)}{P(B)}$

Bayes theorem

Example:

- We have constructed spam filter that
  - identifies spam mail as spam with probability 0.95
  - Identifies usual mail as spam with probability 0.005
- This kind of spam occurs once in 100,000 mails
- If we found that a letter is a spam, what is the probability that it is actually a spam?

Bayes theorem

- We have some knowledge about event B
  - Prior probability P(B) of B
- We get new information A
  - P(A)
  - P(A|B) probability of A can occur given B has occured
- New (updated) knowledge about B
  - Posterior probability P(B|A)

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

Random variables

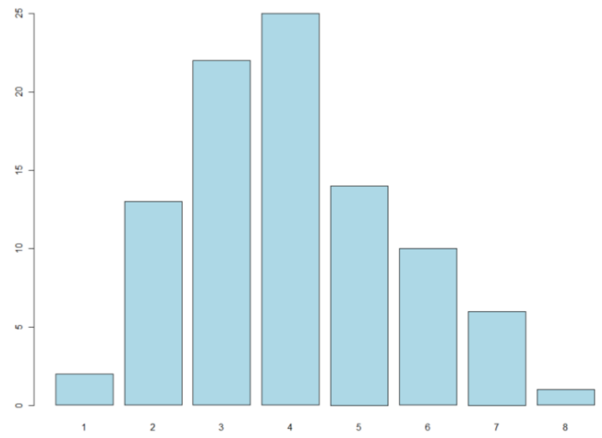
- Instead of having events, we can have a variable X:
  - Events→$\mathbb{R}$ Continuous random variables
  - Events→$\mathbb{N}$ Discrete random variables

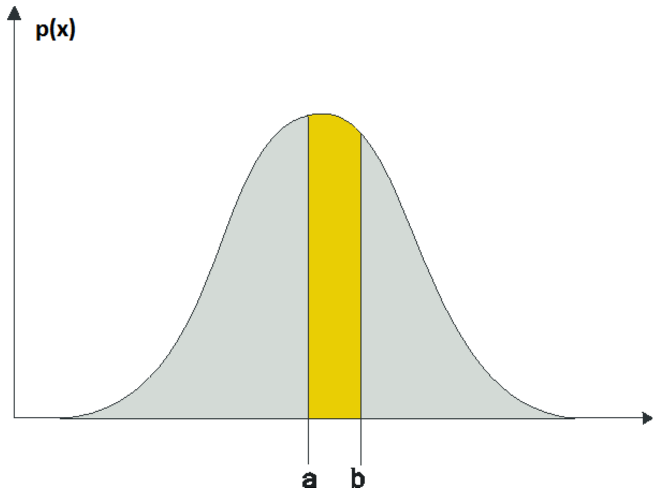
Examples:

- X={amount of times the word ”crisis” can be found in financial documents}
  - P(X=3)
- X={Time to download a specific file to a specific computer}
  - P(X=0.36 min)

Distributions

- Discrete
  - Probability mass function P(x) for all feasible x
- Coninuous
  - Probability density function p(x)
    - $p(x \in [a,b]) = \int_a^b p(x)dx$
    - $p(x) \geq 0, \int_{-\infty}^{+\infty} p(x)dx = 1$
  - Cumulative distribution function $F(x) = \int_0^x p(t)dt$





Expected value and variance

- Expected value = mean value
  - $E(X) = \sum_{i=1}^{n} X_i P(X_i)$
  - $E(X) = \int X p(X) dX$

- Variance how much values of random variable can deviate from mean value
  - $Var(X) = E\big(X - E(X)\big)^2 = E(X^2) - E(X)^2$

Probabilities

- **Laws of probabilities**
  - Sum rule (compute **marginal** probability)

$$p(X) = \sum_Y p(X,Y)$$

$$p(X) = \int p(X,Y)dY$$

  - Product rule

$$p(X,Y) = p(X|Y)p(Y)$$

Combination 1:

$$p(X) = \sum_Y p(X|Y)p(Y)$$

$$p(X) = \int p(X|Y)p(Y)dY$$

Bayes theorem

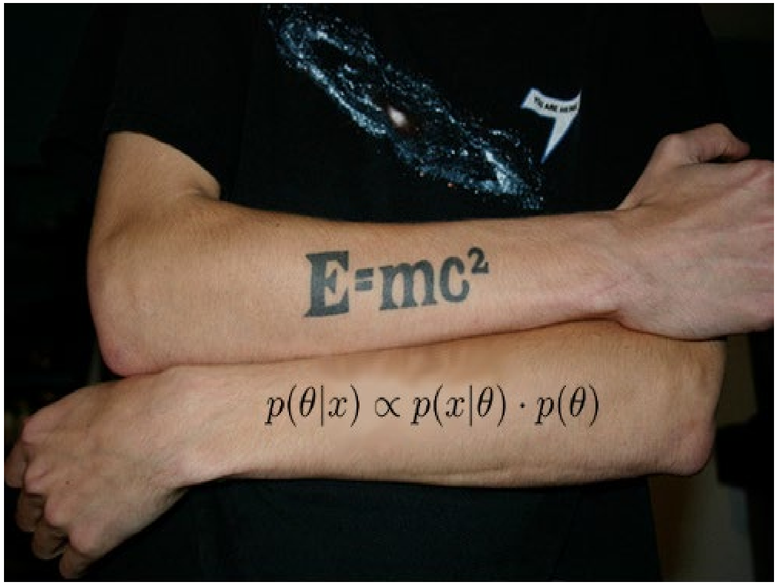
For random variables:

Bayes Theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$$p(Y|X) \propto p(X|Y)p(Y)$$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{\int p(X|Y)p(Y)dY}$$



Some conventional distributions

Bernoulli distribution

– Events: Success (X=1) and Failure (X=0)
– P(X=1)=p, P(X=0)=1-p

– $E(X) = p$
– $Var(X) = p(1-p)$

Examples: Tossing coin, vinning a lottery,..

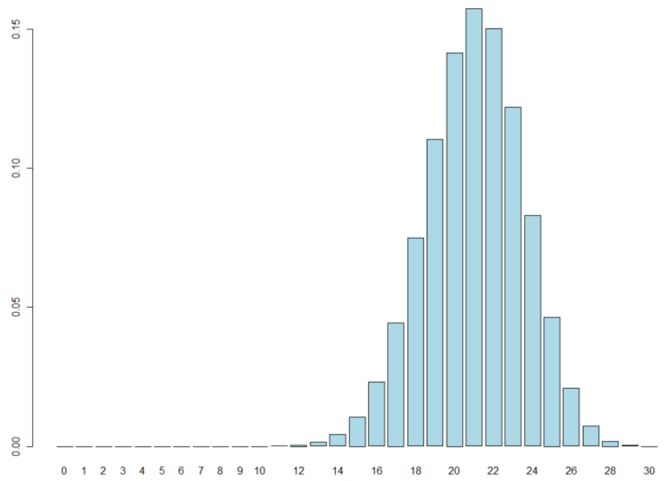
Some conventional distributions

Binomial distribution

- Sequence of *n* Bernoulli events
- X={Amount of successes among these events}, X=0,…,n

$$P(X=r)=\frac{n!}{(n-r)!\,r!}p^r(1-p)^{n-r}$$

- $EX = np$
- $Var(X) = np(1-p)$



Poisson distribution

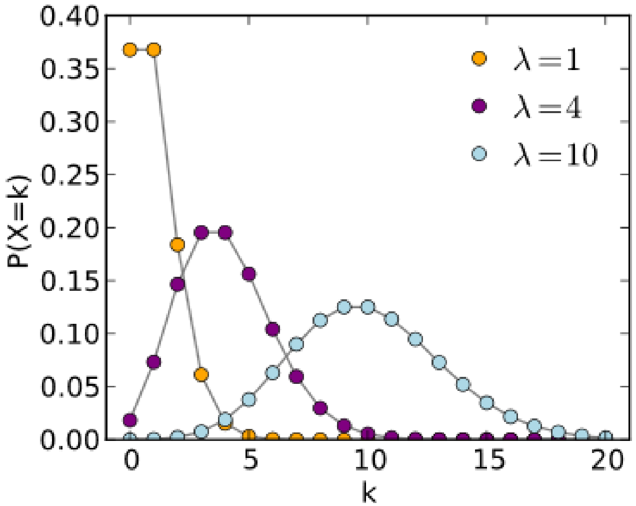
- Customers of a bank **n** (in theory, endless population)
- Probability that a specific person will make a call to the bank between 13.00 and 14.00 a certain day is ***p***
  - ***p*** can be very small if population is large (rare event)
  - Still, some people will make calls between 13.00 and 14.00 that day, and their amount may be quite big
  - A known quantity **λ=*np*** is mean amount of persons that call between 13.00 and 14.00
  - **X**={amount of persons that have called between 13.00 and 14.00}

Poisson distribution

- $P(X = r) = \lim_{n \to \infty} \frac{n!}{(n-r)!r!} p^r (1-p)^{n-r}$
- It can be shown that

$$P(X = r) = \frac{\lambda^r e^{-\lambda}}{r!}$$

- $E(X) = \lambda$
- $Var(X) = \lambda$



Poisson distribution

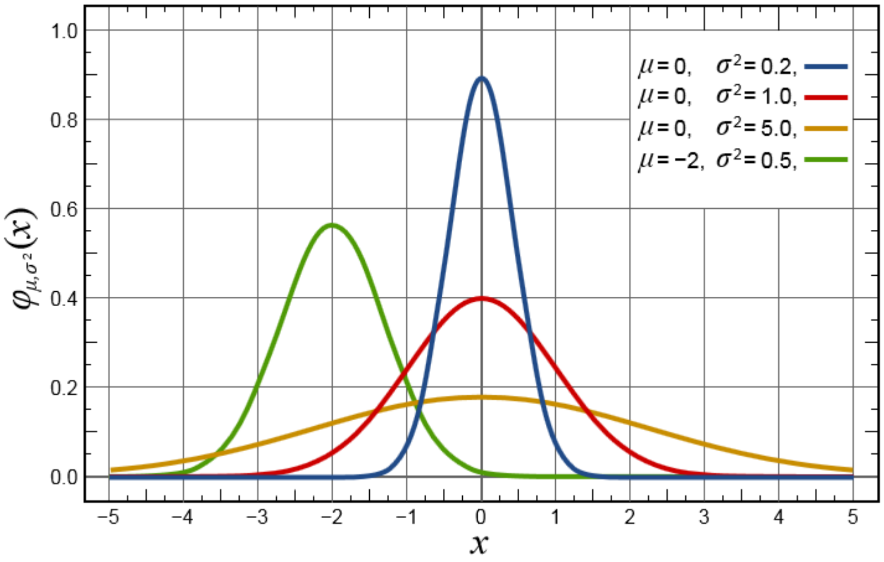
- Further properties:
  - Poisson distribution is a good approximation of the binomial distribution if n >20 and $p < 0.05$
  - Excellent approximation if $n \geq 100$ and $np \leq 10$

Normal distribution

- Appears in almost all applications
  - Difference between the times required to download two specific documents to a specific computer

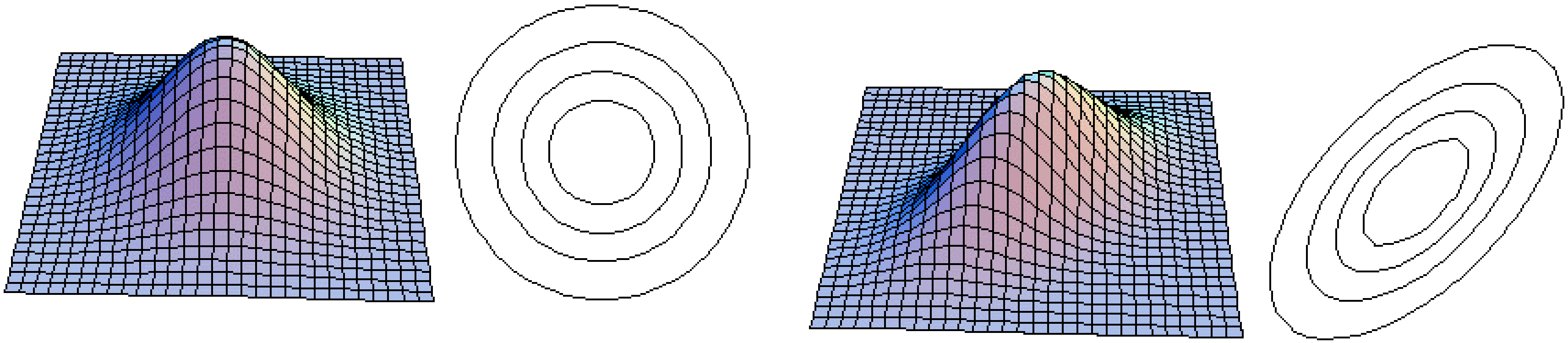
$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \sigma > 0$$

- $E(X) = \mu$
- $Var(X) = \sigma^2$



Multivariate distributions

- Probability of two variables having certain values at the same time
  - P.D.F. p(x,y)
  - Correlation



Basic ML ingridients

- Data D: observations
  - Features $X_1, .. X_p$
  - Targets $Y_1, \dots, Y_r$

Case	X_1	X_2	Y
1			
2			
…			

- Model $P(x|\, w_1, \dots w_k)$ or $P(y|x, w_1, \dots w_k)$
  - Example: Linear regression $p(y|x,w) = N(w_0 + w_1 x, \sigma^2)$

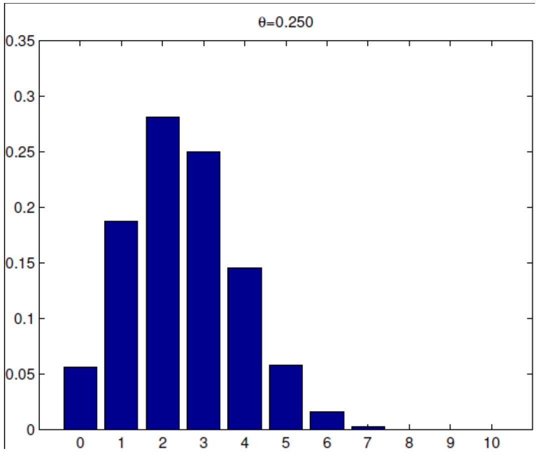
- Learning procedure (data→get parameters $\widehat{w}$ or $p(w|D)$)
  - Maximum likelihood, Bayesian estimation

- Predict new data X^{new} by using the fitted model

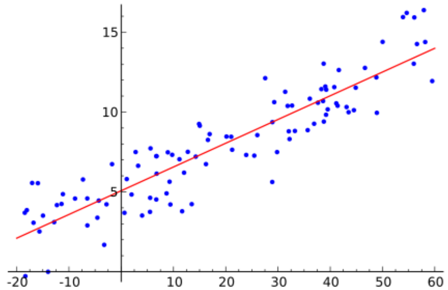
Probabilistic models

- A distribution $p(x|w)$ or $p(y|x,w)$
- Example:
  - $x \sim Bin(n, \theta)$

$$p(x = k|n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$



  - $y \sim N(\alpha_0 + \alpha_1 x, \sigma^2)$



Source: Wikipedia

Learn basic distributions and their properties→PRML, chapter 2!

Fitting a model

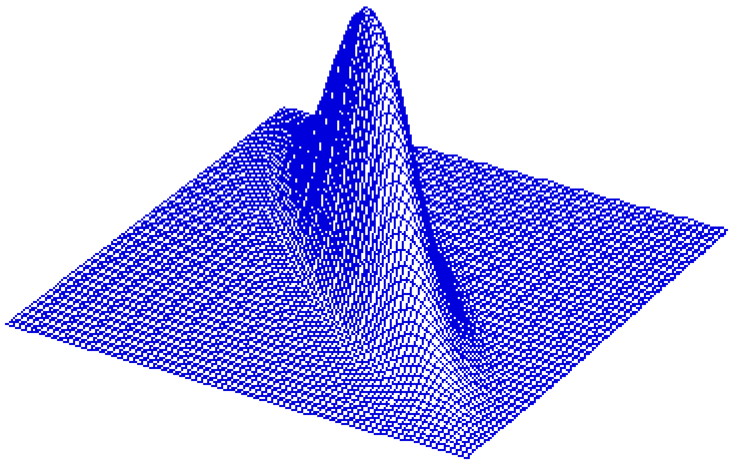
- Given dataset D and model $p(\boldsymbol{x}|w)$ or $p(y|\boldsymbol{x},w)$
  - Frequentist approach: which combination of parameter values fits my data best?
  - Bayesian approach: parameters are random variables, all feasible values are acceptable
    - Different parameter values have different probabilities

Fitting a model

- Frequenist principle: **Maximum likelihood** principle
  - Compute likelihood $p(\boldsymbol{D}|\, w)$

$$p(\boldsymbol{D}|\, w) = \prod_{i=1}^{n} p(X_i|w)$$

$$p(\boldsymbol{D}|\, w) = \prod_{i=1}^{n} p(Y_i|X_i, w)$$



  - Maximize the likelihood and find the optimal w^*

Fitting a model

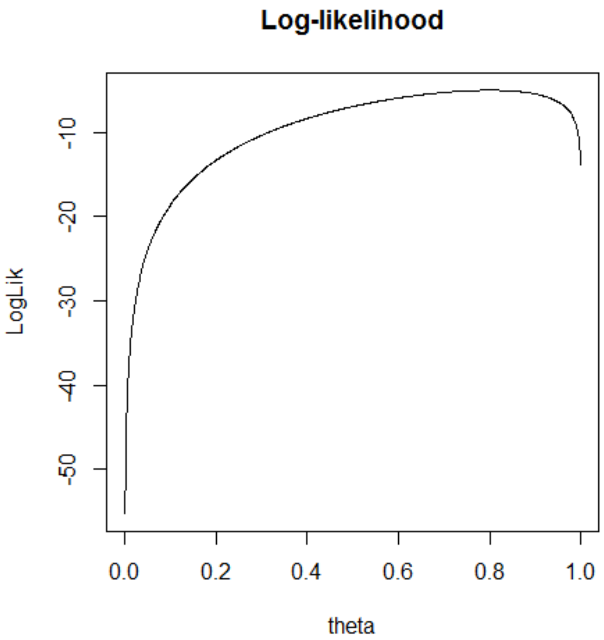
Remarks:

- Likelihood shows how much the chosen parameter value is proper for a specific model and the given data
- Normally **log-likelihood** is used in computations instead
- Other alternatives to ML exist…

Fitting a model

Example: tossing a coin.

$$D = \{0,1,1,0,1,1,1,1,1,1,1,1\},$$
$$p(x=1|\theta)=\theta, p(x=0|\theta)=1-\theta$$



Bayesian probabilities

- Probability reflects your knowledge (uncertainty) about a phenomenon → **subjective probabilities**
  - **Prior probability** $p(w)$, can be uninformative $p(w) \propto 1$
  - Formulate a model, compute **likelihood** $p(D|w)$
  - **Posterior probability** $p(w|D)$, after observing data
    - $p(w|D) \propto p(D|w)p(w)$
- Model parameters are considered as random variables
  - In real life, do not need to be random, but we model as random

Fitting a model

- Bayesian principle
  - Compute $p(w|D)$and then decide yourself what to do with this (for ex. MAP, mean, median)
- Use bayes theorem

$$\boxed{p(w|D) = \frac{p(D|w)p(w)}{p(D)} \propto p(D|w)p(w)}$$

- $p(D)$ is **marginal likelihood**
  - $p(D) = \int p(D|w)p(w)dw$ or
  - $p(D) = \sum_i p(D|w_i)p(w_i)$

Example: tossing a coin. Find $p(\theta|D)$, estimate posterior mean θ^*

Fitting a model

- How to chose the prior?
  - Expert knowledge about the phenomenon
  - Forcing a model to have a certain structure
    - Example: decision trees: prior prefers smaller trees

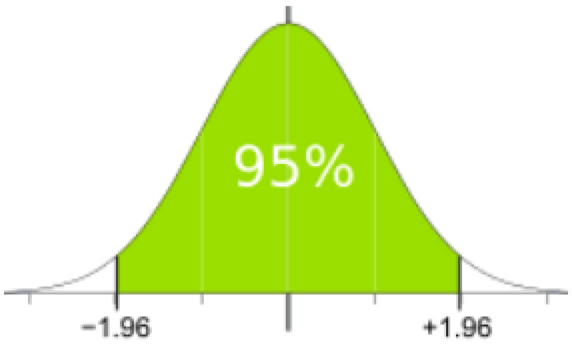
http://en.wikipedia.org/wiki/Conjugate_prior

  - Conjugacy
    - Distribution of the posterior is the same type as the distribution of the likelihood or prior

- Prior is the most controversial about Bayesian methods, but
  - When $N \to \infty$, data overwhelms the prior

Measuring uncertainty

- **Confidence interval** (frequentist)
  1. Model $p(x|w)$ is known
  2. $\hat{w}$ is a function of x by ML
  3. Derive distribution of $\hat{w}$
  4. Compute quantiles
- **Credible interval** (Bayes)
- **Prediction interval** (models)



- Example: Prediction interval for $Y \sim N(2x+4, 1)$ at $x = 5$