Machine Learning I
Week 1: Introduction and Prerequisites

Christian Osendorfer, Martin Felder

Technische Universität München

2009/10/22

What is Machine Learning?

Wikipedia says ...

> *Machine learning [...] is concerned with [...] techniques that allow computers to "learn".*
>
> *Inductive machine learning methods extract rules and patterns out of massive data sets.*
>
> *The major focus of machine learning research is to extract information from data automatically, by computational and statistical methods.*

Tom Mitchell[1]

> *How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes.*

Ideas from: *Statistics, Computer Science*, Physics Computational Neuroscience, Engineering, ...

[1] The Discipline of Machine Learning

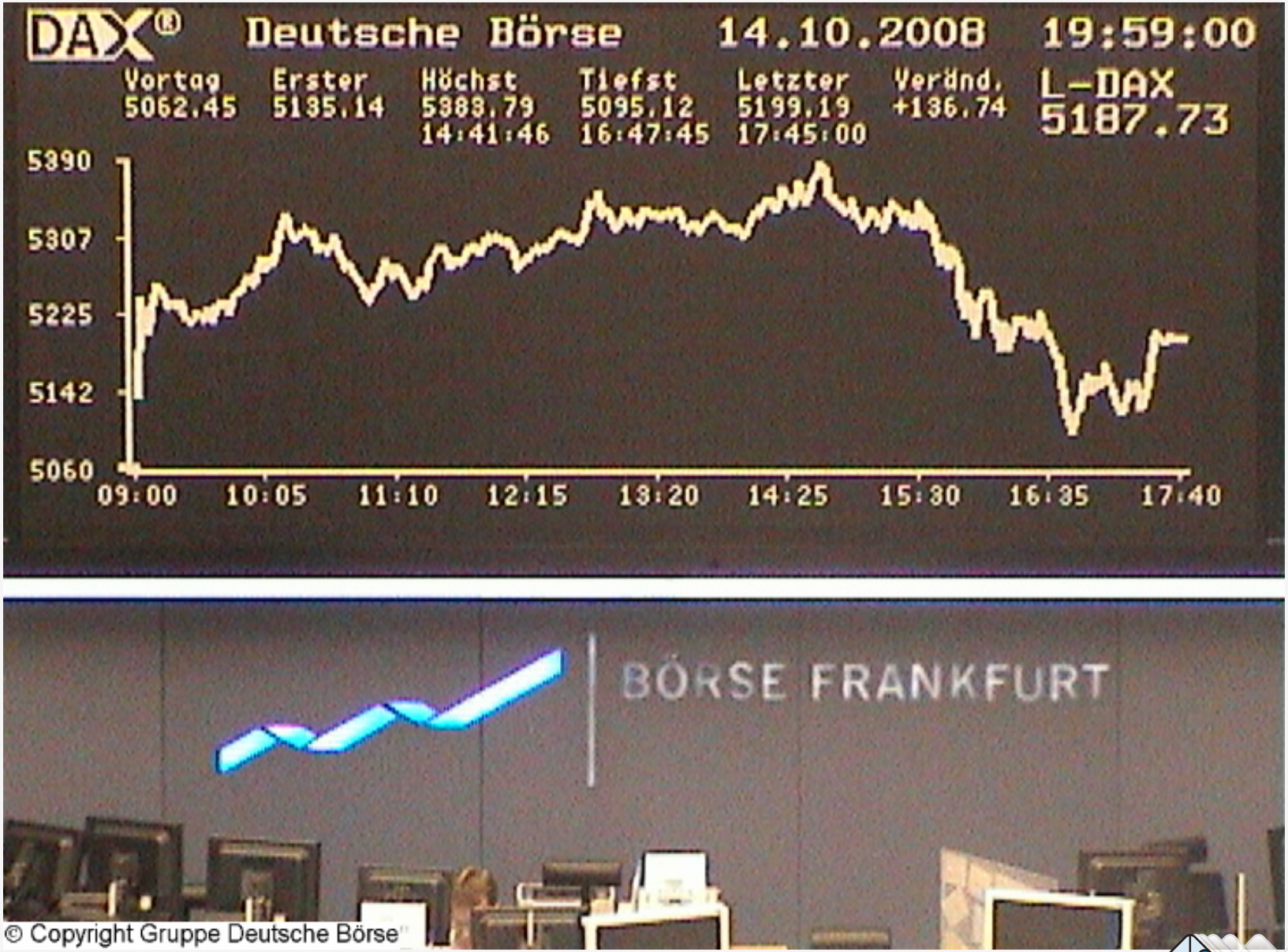
Machine "Learning"

> *Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the task or tasks drawn from the same population more efficiently and more effectively the next time.*
>
> *Herbert Simon*

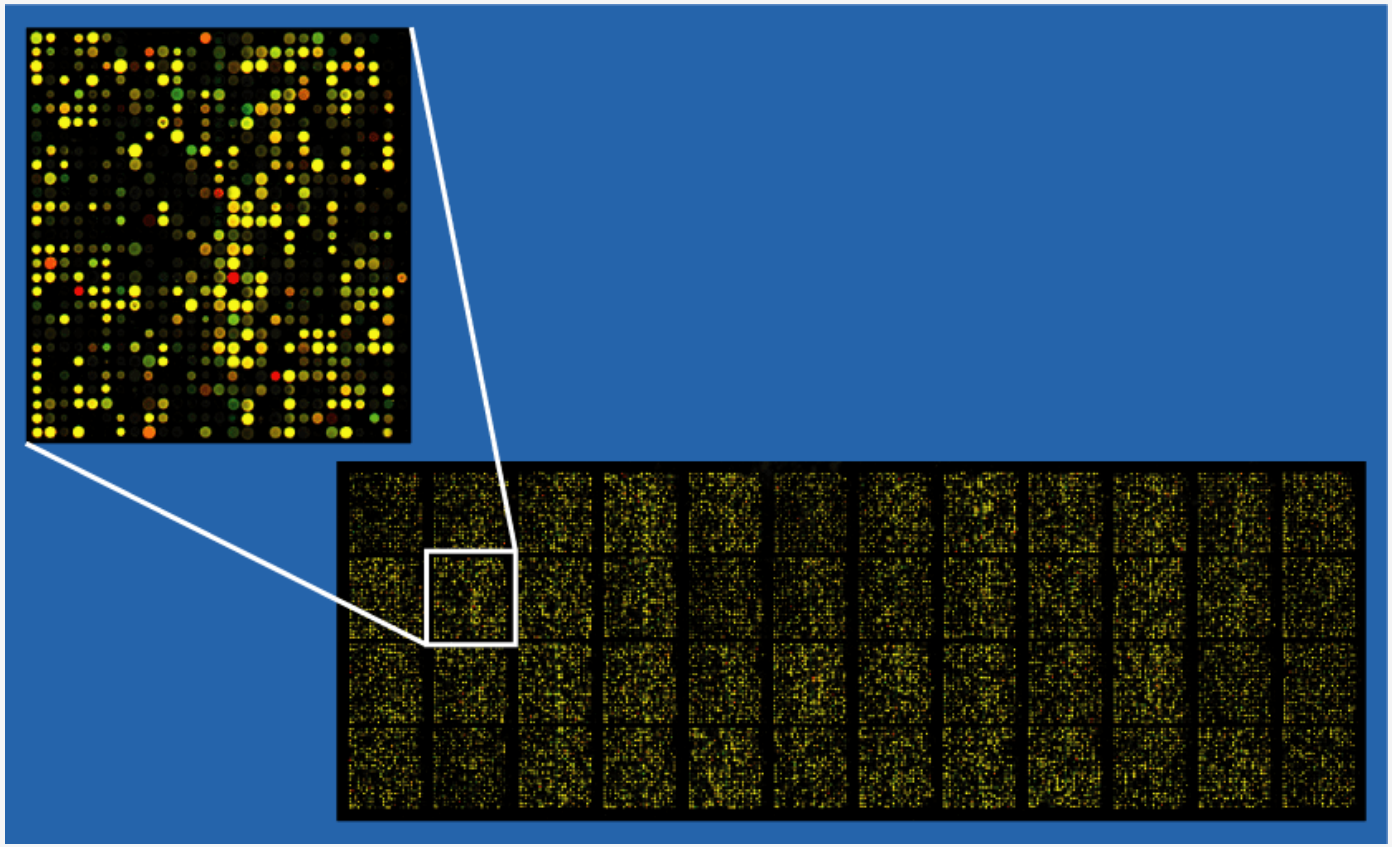
No one needs "learning" to compute a minimum spanning tree in a graph.

Machine Learning is programming computers to optimize a performance criterion using example data or past experience.

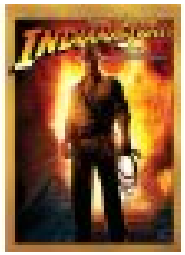
Stock Market Prediction



DNA Microarray Analysis



Collaborative Filtering



General information, tutorials, . . .

- Time and date(s)?!
- Books ?!
- Tutorials will be every Friday morning in Room MI 00.08.038.
- Discuss *assignments*.
- Assignments: Out on Thursday, hand in one week later.
- Collaboration: (up to) 3 people can hand in one assignment together.
- In order to take the final exam, you need 2/3 of the assignments – if you have problems with that, let us know *right now*.

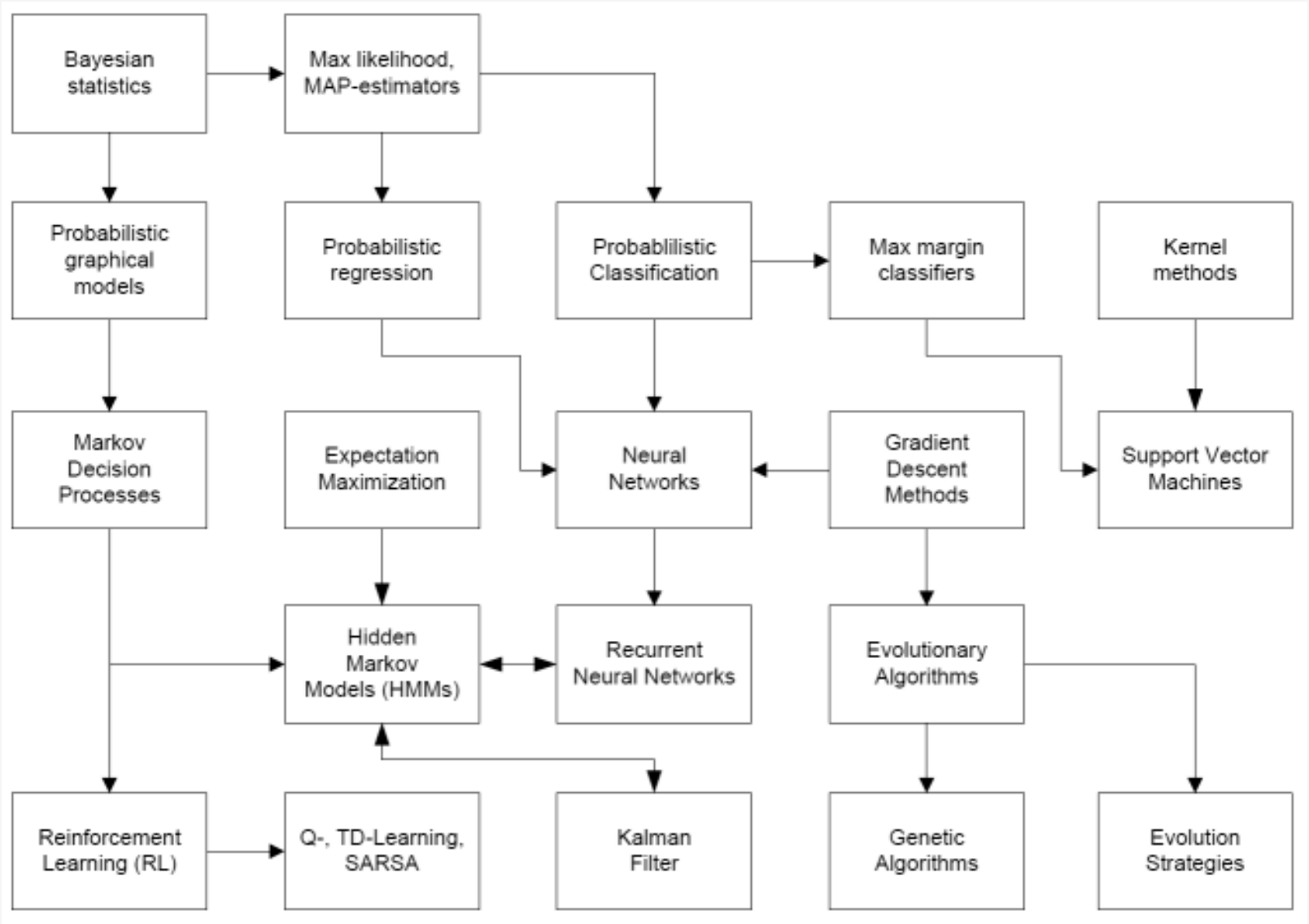
(Canonical) types of ML

Supervised Learning For a given input, the learner is also provided with the desired output. The goal is to learn to produce correct outputs for unseen inputs.

Unsupervised Learning The goal is to build a model of the inputs (e.g. for clustering, outlier detection, compression, ...), without knowing in advance what to actually look for.

Reinforcement Learning Instead of simple outputs, the learner produces actions that affect the state of the world. Depending on these actions, the learner receives rewards. The goal is to learn to act in a way that maximizes rewards in the long term.

Structure Overview

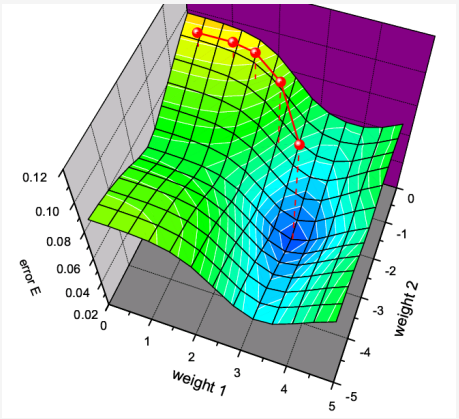


Linear Algebra

- Vector space, linear independence, (orthogonal) basis, (symmetric) matrices, Determinant, Eigenvectors, Eigenvalues.
- Any real symmetric matrix can be diagonalized.
- QR decomposition of a real symmetric matrix A: $A = QR$, with $Q^T Q = QQ^T = I$ and R an upper triangular matrix.
- Singular Value Decomposition (SVD): Factorization of a *rectangular* real or complex matrix. Suppose $A \in \mathbb{R}^{m \times n}$.
  Then
  $$A = U \Sigma V^T$$
  with $U \in \mathbb{R}^{m \times m}$, $U^T U = UU^T = I$, Σ is a diagonal matrix with non negative entries and $V \in \mathbb{R}^{n \times n}$, $V^T V = VV^T = I$.
- Use QR/SVD to solve[2] system of linear equations $Ax = y$.

[2] *I must also confess a strong bias against the fashion for reusable code. To me "re-editable code" is much, much better than an untouchable black box or toolkit.* Donald Knuth

Calculus



- The gradient $\nabla g(a)$ of a given function $g : \mathbb{R}^n \mapsto \mathbb{R}$ points in the direction of the greatest rate of increase of g in a.
- Chain rule: $(g \circ h)'(x) = g'(h(x))h'(x)$.

- Hessian Matrix: second derivatives.
- The method of Lagrange multipliers: Find the local extrema of a function subject to one or more constraints.
- Jensen's inequality: For a real, convex function g defined on interval $\mathcal{I}$ and positive weights $\lambda_1, \ldots, \lambda_n$ with $\sum_{i=1}^{n} \lambda_i = 1$ the following holds:

$$g\left(\sum \lambda_i x_i\right) \leq \sum \lambda_i g(x_i)$$

for any $x_1, \ldots, x_n \in I$.

Probability Basics

You already know a lot about probability theory, e.g.

- Kolmogorov Axioms
- (discrete/continuous) Random Variables
- Expectation
- Variance
- Independence
- Conditional Probability
- i.i.d.

Some basic rules that are important for Machine Learning:

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

Product rule $p(X, Y) = p(Y|X) \cdot p(X)$

Bayes' rule $p(Y|X) = \frac{p(X|Y) \cdot p(Y)}{p(X)}$

(X, Y are random variables)

Conditional Independence

Conditional Independence is one of the most basic assumptions taken to make problems computationally tractable.

- X and Y are conditionally independent given Z iff $p(X|Y,Z) = p(X|Z)$.
- I.e. Y does not provide any information about X if Z is already known.
- This is written as: $X \perp Y|Z$.
- Intuition tells us that $X \perp Y|Z \Leftrightarrow Y \perp X|Z$.
- The joint conditional probability decomposes: $X \perp Y|Z \Leftrightarrow p(X,Y|Z) = p(X|Z)p(Y|Z)$

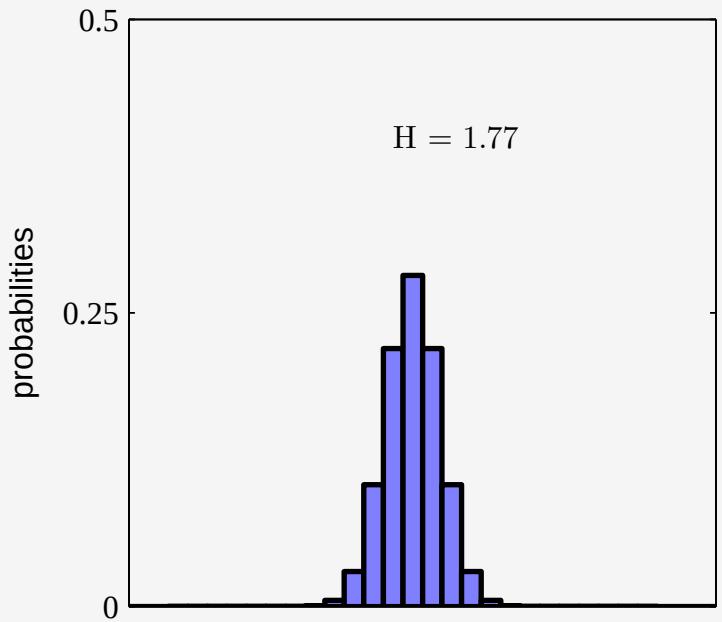
Example application: Naïve Bayes Classifier

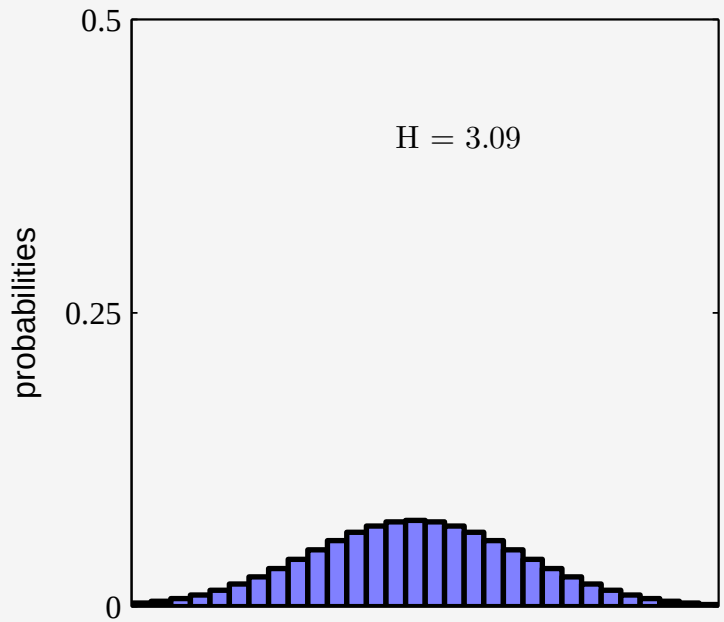
Information Theory

Shannon Entropy, or information entropy, is a measure for the information contained in a random variable (or message, picture, ...). In the discrete case:

$$\mathrm{H}[x] = -\sum_x p(x) \log_b p(x)$$

The unit of H is *bit* for $b = 2$, *nat* for $b = \mathrm{e}$, and *dit* (or digit) for $b = 10$. By default we will use $b = \mathrm{e}$.





Information Theory

Entropy in the continuous case = differential entropy

$$\mathrm{H}[x] = -\int p(x) \ln p(x) \, \mathrm{d}x$$

What about maximum differential entropy?

- there is no uniform distribution in the unbounded continuous case!
- turns out Gaussian has maximum entropy:

$$\mathrm{H}[x] = \frac{1}{2}(1 + \ln(2\pi\sigma^2))$$

- Conditional entropy of y given x:

$$\mathrm{H}[y|x] = -\int \int p(y, x) \ln p(y|x) \, \mathrm{d}x \, \mathrm{d}y$$

Comparing Distributions

Assume we have a true distribution $p(x)$, and model it by $q(x)$. How good is our model?
The Kullback-Leibler divergence is a measure for the similarity of distributions:

$$\mathrm{KL}(p||q) = -\sum_x p(x) \ln \frac{q(x)}{p(x)}$$

Note that $\mathrm{KL}(p||q) \geq 0$, but KL it is *not* a distance metric, because $\mathrm{KL}(p||q) \neq \mathrm{KL}(q||p)$.
A related measure is Mutual information:

$$\mathrm{I}[x, y] = \mathrm{KL}(p(x, y)||p(x)p(y))$$

which can also be written as

$$\begin{aligned} \mathrm{I}[x, y] &= \mathrm{H}[x] - \mathrm{H}[x|y] && (1) \\ &= \mathrm{H}[y] - \mathrm{H}[y|x] && (2) \\ &= \mathrm{H}[x] + \mathrm{H}[y] - \mathrm{H}[x, y] && (3) \end{aligned}$$

Probability Distributions

The Bernoulli distribution takes value 1 with success probability θ and value 0 otherwise:

$$\text{Bern}(x|\theta) = \begin{cases} \theta & \text{if } x = 1, \\ 1-\theta & \text{if } x = 0, \\ 0 & \text{otherwise.} \end{cases}$$

- special case of the Binomial distribution with $N = 1$ trials
- hence can also be written as

$$\text{Bern}(x|\theta) = \theta^x (1-\theta)^{1-x}$$

- we'll get back to this later!

Probability Distributions

- Beta:

$$\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}, \quad \mu \in [0,1]$$

- Multivariate Gaussian distribution:

$$N(\mathbf{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- also Gamma, Wishart, Dirichlet, Von Mises, Student's t, ...

Beta Distribution

Beta is a distribution over $\mu \in [0, 1]$

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

$$\mathbb{E}[\mu] = \frac{a}{a+b}$$

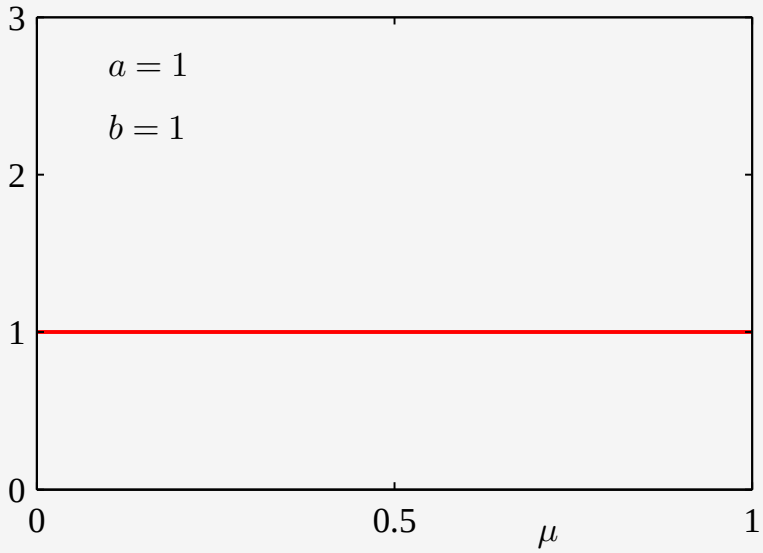
$$\text{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$$

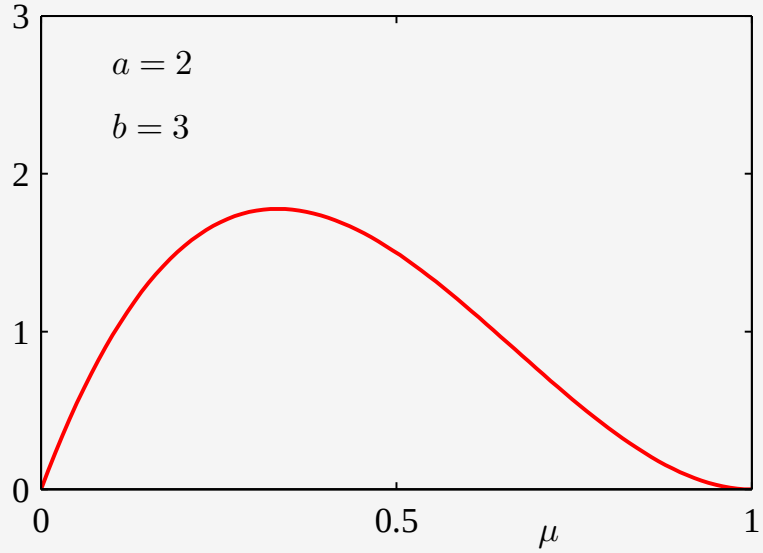
The prefactor is for normalization, with $\Gamma(x) = \int_0^\infty e^{-u}u^{x-1}du$ being the generalized factorial function.

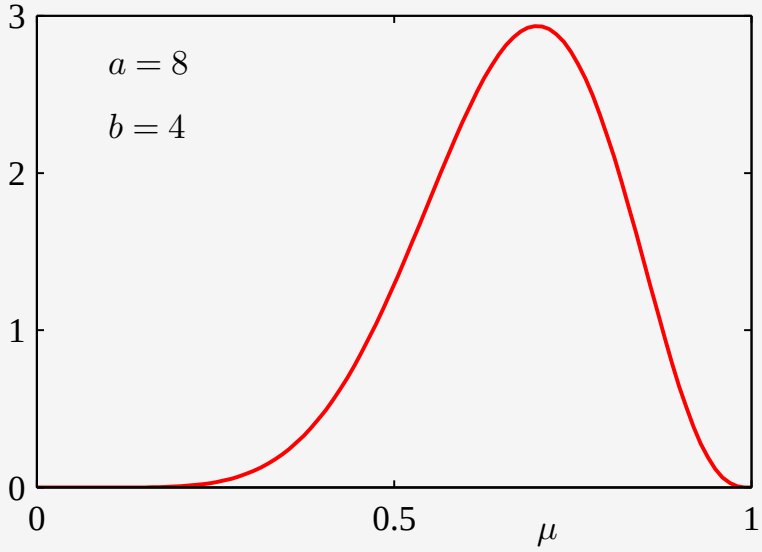
a and b are called hyperparameters of the distribution.

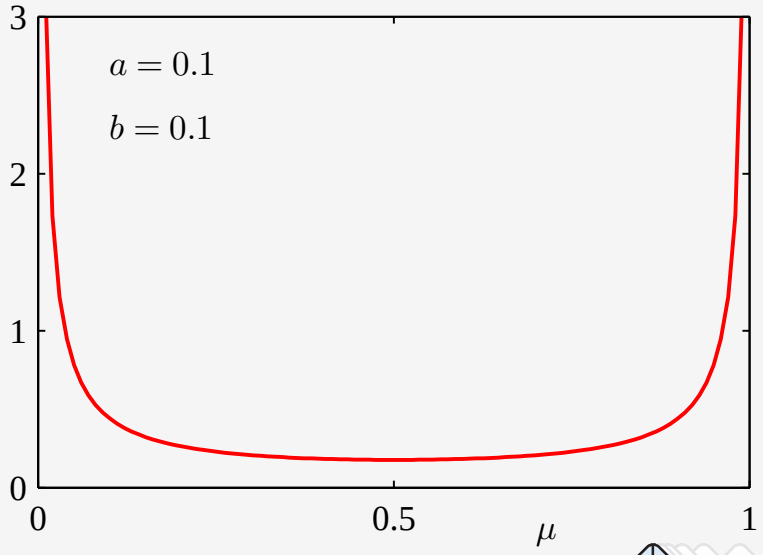
The Beta distribution is the *conjugate prior* of the Bernoulli distribution – we'll get back to this in a minute.

Beta Distribution



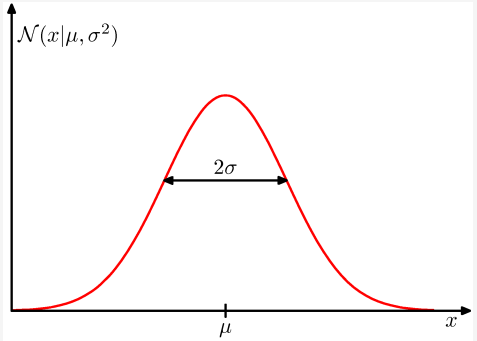






Multivariate Gaussian Distribution

You should know the univariate Gaussian Distribution ...

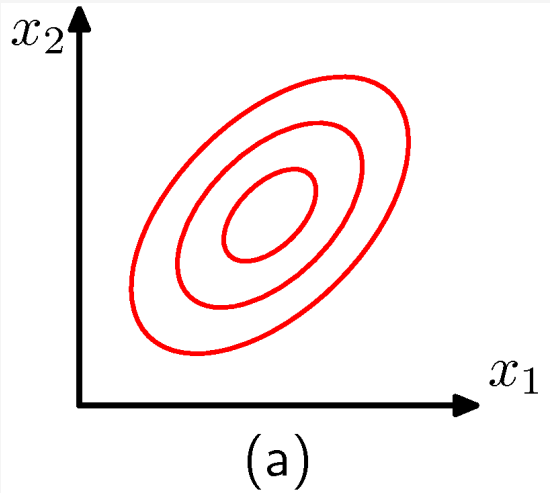


The multivariate Gaussian is the extension of the univariate Gaussian to n dimensions.

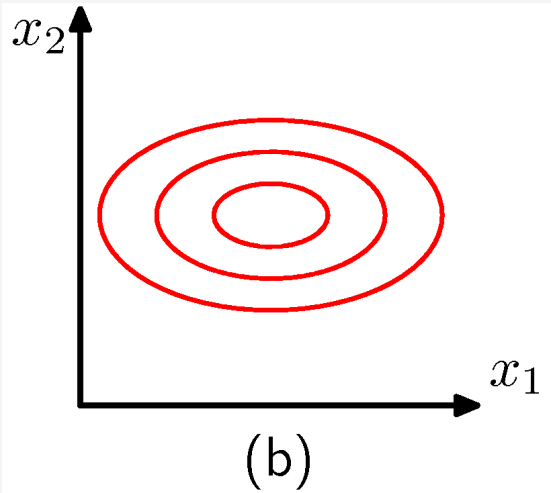
- $N(\mathbf{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$
- Σ is symmetric matrix w.l.o.g! It is called the covariance matrix.
- One can show with some fancy linear algebra (diagonalization) that this is actually a distribution (i.e. it is normalised).
  Even though it looks complicated, everything is as *expected*:
- $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$
- $\text{cov}[\mathbf{x}] = \Sigma$

Multivariate Gaussian Distribution

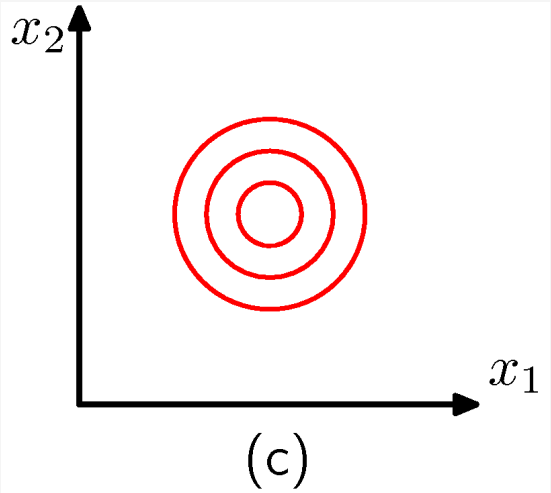
The covariance matrix is defined as: $\text{cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T]$



general Σ



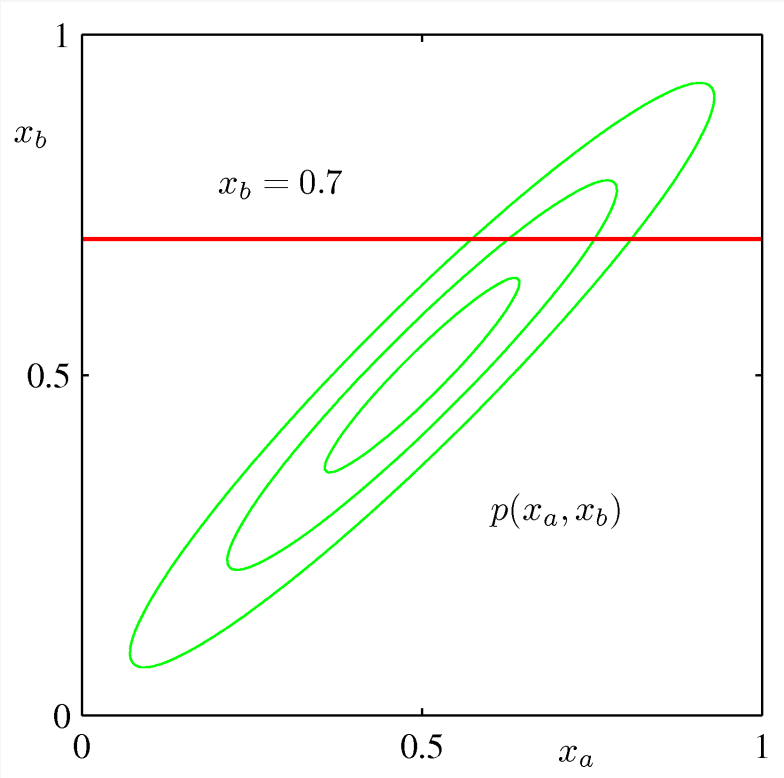
Σ diagonal

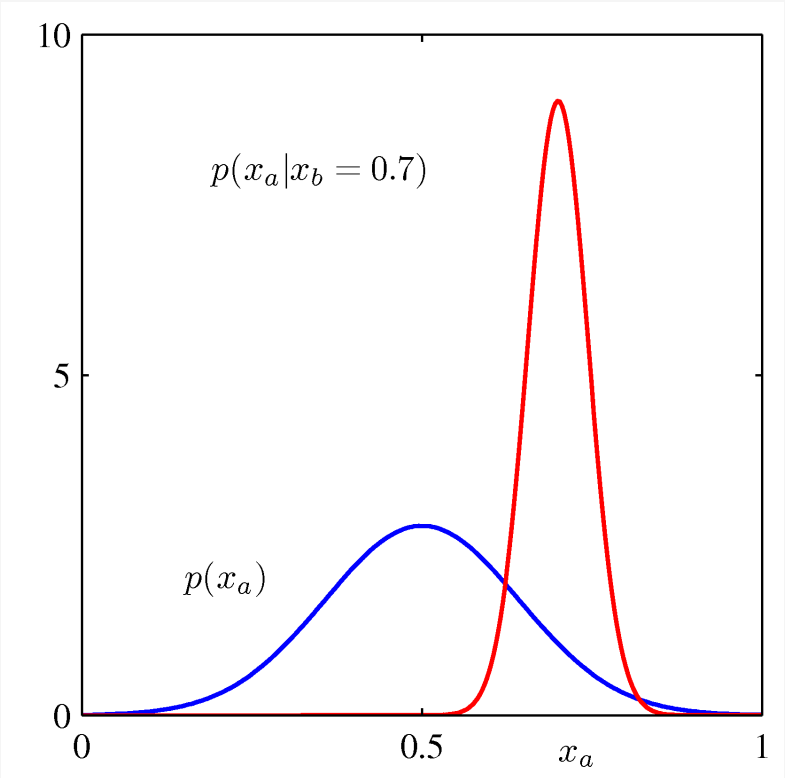


$\Sigma = \sigma \mathbf{I}$

Multivariate Gaussian Distribution

Conditional and marginal distributions are also Multivariate Gaussian!





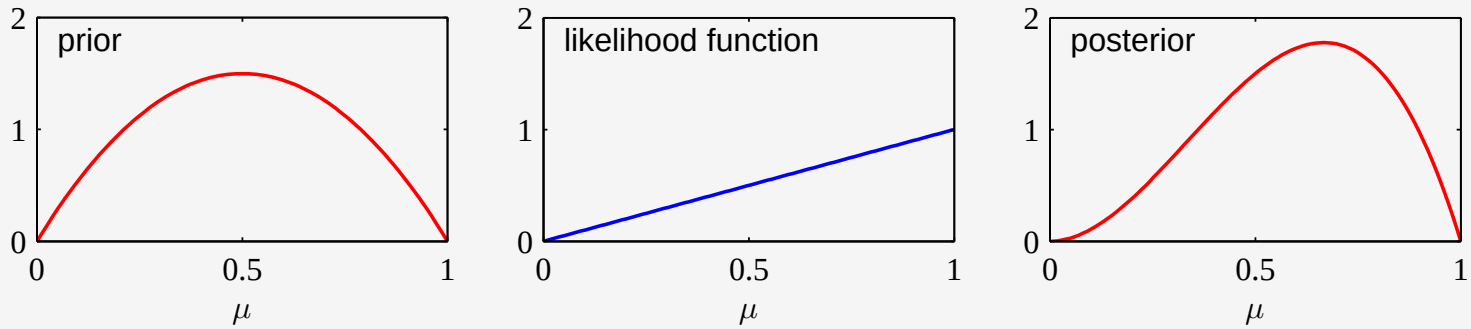
"My first machine learning problem"

We are now ready to develop our first ML algorithm.

- recap from introduction: want to extract rules/patterns from data
- what does this mean in a concrete case?
- obviously data-driven approach, inductive
- simple example: coin flip – can generate observations easily: $x_1 = \text{head}, x_2 = \text{head}, x_3 = \text{tail}, \ldots$
- find the *rule* that explains observations $\mathcal{D} = x_1, x_2, \ldots$ *best*
- the rule can take the form of some model, which has parameter θ
- the likelihood of our observations for a given rule θ is $p(x_1, x_2, \ldots | \theta)$
- how do we find the *best* rule θ? $\rightarrow$ yields maximum likelihood!
- but joint distribution of all observations is hard to deal with
- make i.i.d. assumption! $p(x_1, x_2, \ldots | \theta) = \prod p(x_i | \theta)$
- introduce model assumption: $p(x_i | \theta)$ is probably a Bernoulli distribution

- N tosses of a coin, say we get m heads (assume i.i.d. tosses)
- what is the probability for the next toss to yield 'head'?
- trick: Log likelihood (MLE)
  $\ell(\theta) = \ln p(\mathcal{D}|\theta) \overset{(i.i.d.)}{=} \ln \prod p(x_i|\theta) = \sum \ln p(x_i|\theta)$
- log is convex, does not change location of maximum
- differentiate wrt. θ, and set equal zero
- $p(x = \text{head}) = \text{argmax}_\theta \, \ell(\theta) = \theta_{ML} = m/N$!
- MLE is consistent: $\lim_{N \to \text{inf}} \theta_{ML} = \theta$
- Problem: Point estimate! MLE is often a bad idea: e.g. $N = 3$, got three heads $\Longrightarrow p(x = \text{head}) = 1$?!
- intuition says next 3 trials may well produce a tail
- but if $N = 3000$, got 3000 heads: our intuition is not worth much
- need to incorporate a prior belief to modulate results of a small number of trials
- θ itself then has some distribution

- calculate the posterior distribution from Bayes' law: $p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$
- very important idea! In words: posterior $\propto$ likelihood $\times$ prior



- $\propto$ is enough, because can always normalize to $\int p \mathrm{d}\theta = 1$
- ideally use a conjugate prior: given $p(\mathcal{D}|\theta)$ (here: Bernoulli), choose distribution $p(\theta)$ such that $p(\theta|\mathcal{D})$ ends up having the same functional form
- because then can use it as a prior for the next experiment!
- prior for Bernoulli is the Beta distribution, with (hyper-)parameters a, b
- hence our prior is $p(\theta|a, b) = \mathrm{Beta}(\theta|a, b)$

this yields the posterior

$$\begin{aligned} p(\theta|\mathcal{D}, a, b) &\propto p(\mathcal{D}|\theta)p(\theta|a, b) \\ &= \theta^m(1-\theta)^{N-m} \cdot \theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{m+a-1}(1-\theta)^{N-m+b-1} \end{aligned}$$

- $a-1$ and $b-1$ can be interpreted as *previous* heads and tails!
- hence now have a solid statistical model for θ, where observations have been assimilated into the prior
- e.g. $a = b = 5$, $N = 3$ with three heads coming up: $p(\text{head}) = \theta_{MAP} = \text{argmax}_\theta\, p(\theta|\mathcal{D}, a, b) = \frac{7}{11}$
- this is called maximum a posteriori (MAP) estimate
- but wait a minute, we can do better than that: Given the hyperparameters, we can calculate the full predictive distribution

$$
\begin{aligned}
p(x|D,a,b) &= \int_0^1 p(x,\theta|D,a,b)\,\mathrm{d}\theta \\
&= \int_0^1 p(x|\theta)p(\theta|D,a,b)\,\mathrm{d}\theta \qquad \text{(used cond. indep.!)} \\
&= \int_0^1 \theta^x(1-\theta)^{1-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\,\mathrm{d}\theta \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 \theta^{x+a-1}(1-\theta)^{b-x}\,\mathrm{d}\theta \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(x+a)\Gamma(b-x+1)}{\Gamma(a+b+1)}
\end{aligned}
$$

- now remember that $\Gamma(a) = (a-1)!$ for $a \in \mathbb{N}$
- using the example above, with the three additional heads assimilated, we have $a = 8, b = 5$
- inserting this for $x = 1$ yields $p(x = 1|a,b) = 8/13$
- this is called the fully Bayesian treatment

Conclusion

We have obtained three different results for the coin flip experiment with $N = 3, m = 3, a = b = 5$, depending on where we collapse our probability distribution and what kind of prior knowledge we introduce:

MLE: $p(x = \text{head}) = 1$
pro: easy to calculate, consistent (large N!)
con: point estimate, misleading for small N

MAP: $p(x = \text{head}) = 7/11 \simeq 0.636$
pro: introduce prior, often still tractable
con: may still fail if posterior multimodal etc.

fully Bayesian: $p(x = \text{head}) = 8/13 \simeq 0.615$
pro: make the most out of your model!
con: usually analytically intractable