Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Mathematical Basics (Bayesian Learning)

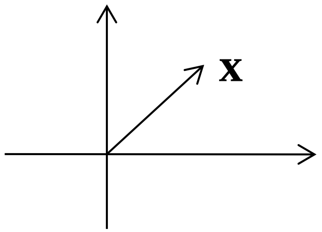
Tobias Scheffer
Uwe Dick

Overview

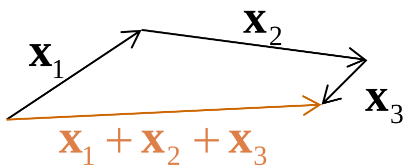
- Linear Algebra:
  - Vectors, Matrices, …
- Analysis & Optimization:
  - Norms, convex functions
- Bayesian statistics
  - Bayesian Learning

Linear Algebra
Vectors

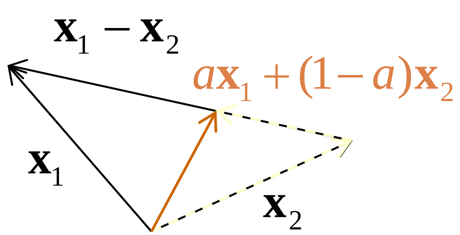
- Vector: $$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} = [x_1 \quad \cdots \quad x_m]^{\mathrm{T}}$$



- Sum of vectors: $$\sum_{i=1}^{n} \mathbf{x}_i = \begin{bmatrix} x_{11} + \ldots + x_{n1} \\ \vdots \\ x_{1m} + \ldots + x_{nm} \end{bmatrix}$$



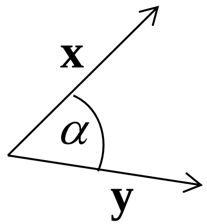
  - Weighted average $$a\mathbf{x}_1 + (1-a)\mathbf{x}_2 = \mathbf{x}_2 + a(\mathbf{x}_1 - \mathbf{x}_2)$$



- Dot product (scalar product / inner product)

$$\langle \mathbf{y}, \mathbf{x} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^{\mathrm{T}}\mathbf{y} = \sum_{i=1}^{m} x_i y_i$$

$$\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\| \|\mathbf{y}\| \cos\alpha$$



Linear Algebra
Matrices

- Matrix: $$\mathbf{X}=\begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}=\begin{bmatrix} x_{11} & \cdots & x_{m1} \\ \vdots & \ddots & \vdots \\ x_{1n} & \cdots & x_{mn} \end{bmatrix}^{\mathrm{T}}=[\mathbf{x}_1 \quad \cdots \quad \mathbf{x}_n]=\begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_m^{\mathrm{T}} \end{bmatrix}$$
- Sum of matrices: $$\mathbf{X}+\mathbf{Y}=[\mathbf{x}_1+\mathbf{y}_1 \quad \cdots \quad \mathbf{x}_n+\mathbf{y}_n]=\begin{bmatrix} \mathbf{x}_1^{\mathrm{T}}+\mathbf{y}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_m^{\mathrm{T}}+\mathbf{y}_m^{\mathrm{T}} \end{bmatrix}$$
- Matrix product: $$\mathbf{YX}\neq\mathbf{XY}=\begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_m^{\mathrm{T}} \end{bmatrix}[\mathbf{y}_1 \quad \cdots \quad \mathbf{y}_n]=\begin{bmatrix} \langle\mathbf{x}_1,\mathbf{y}_1\rangle & \cdots & \langle\mathbf{x}_1,\mathbf{y}_n\rangle \\ \vdots & \ddots & \vdots \\ \langle\mathbf{x}_m,\mathbf{y}_1\rangle & \cdots & \langle\mathbf{x}_m,\mathbf{y}_n\rangle \end{bmatrix}$$

Linear Algebra
Matrices

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

- Quadratic: $n = m$
- Symmetric: $\mathbf{A} = \mathbf{A}^{\mathrm{T}}$
- Positive definite: $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} > 0 \quad \forall \mathbf{x} \neq \mathbf{0}$ if $\mathbf{A}$ symmetric
- trace: $tr(\mathbf{A}) = \sum_{i=1}^{m} a_{ii}$
- rank: $rk(\mathbf{A}) = \#\text{linearly independent rows/columns}$

Linear Algebra
Special Matrices

- Vector / Matrix of all ones:

$$\mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \mathbf{1} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$$

- Unit vector:

$$\mathbf{e}_i = [\underbrace{0 \quad \cdots \quad 0}_{i-1} \quad 1 \quad 0 \quad \cdots \quad 0]^{\mathrm{T}}$$

- Diagonal matrix:

$$diag(\mathbf{a}) = [a_1\mathbf{e}_1 \quad \cdots \quad a_m\mathbf{e}_m] = \begin{bmatrix} a_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_m \end{bmatrix}$$

- Matrix-vector product:

$$\mathbf{X}\mathbf{y} = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_m^{\mathrm{T}} \end{bmatrix} \mathbf{y} = \begin{bmatrix} \langle \mathbf{x}_1, \mathbf{y} \rangle \\ \vdots \\ \langle \mathbf{x}_m, \mathbf{y} \rangle \end{bmatrix}$$

Linear Algebra
Distances and Norms

- Examples for vector distances and norms:
  - *p*-norm: $\|\mathbf{x}\|_p = \sqrt[p]{\sum_{i=1}^{m} |x_i|^p}$
  - Manhattan norm: $\|\mathbf{x}\|_1$
  - Euclidian norm: $\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\sum_{i=1}^{m} x_i^2}$

Distance between $\mathbf{x}$ and $\mathbf{y}$:

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$$

- Examples of matrix norms:
  - *p*-norm $\|\mathbf{X}\| = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |x|_{ij}^p \right)^{\frac{1}{p}}$
  - Frobenius norm: $\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}^2}$

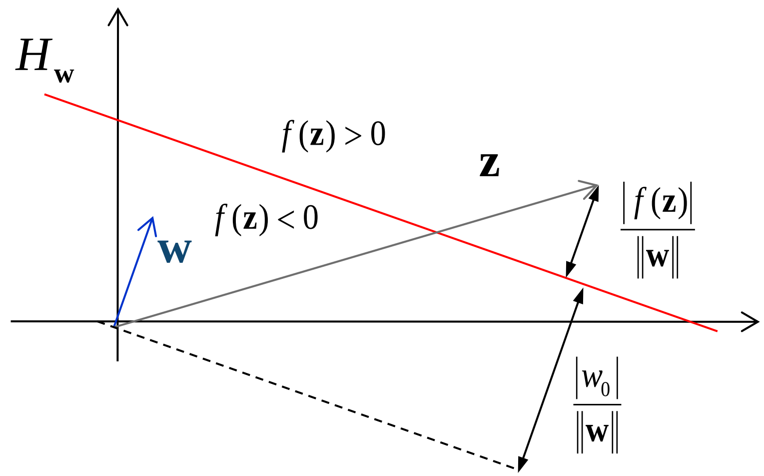
Distance between $\mathbf{X}$ and $\mathbf{Y}$:

$$d(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\|$$

Linear Algebra
Geometry

- Hyperplane:

$$H_{\mathbf{w}} = \{\mathbf{x} \mid f(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{w} + w_0 = 0\}$$



- Mahalanobis distance (w.r.t. covariance matrix $\mathbf{A} > 0$):

$$d_{\mathbf{A}}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^{\mathrm{T}} \mathbf{A}^{-1} (\mathbf{x} - \mathbf{y})}$$

Scheffer/Dick: Language Technology

Linear Algebra
Representations & Operations

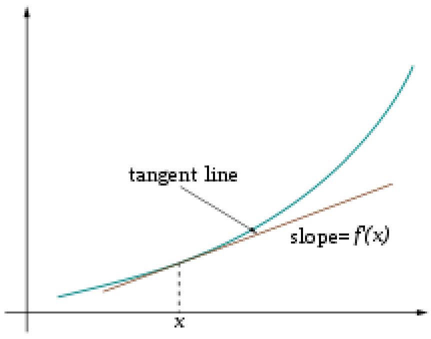
- Representation of data
  - Instance with *m* features: $\mathbf{x} = [x_1, \ldots, x_m]^{\mathrm{T}}$
  - *n* instances (data matrix): $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$
- Decision values (linear function)
  - of a point: $f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$
  - of a data matrix: $f(\mathbf{X}) = \mathbf{w}^{\mathrm{T}}\mathbf{X} + w_0\mathbf{1}$
- Affine-linear transformations of data from $\mathbb{R}^{m_1}$ to $\mathbb{R}^{m_2}$:
  - of a point : $A(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$
  - of a data matrix : $A(\mathbf{X}) = \mathbf{A}\mathbf{X} + \mathbf{B}$
  - Results in reduction of features if $m_2 < m_1$

$\mathbf{A} \in \mathbb{R}^{m_2 \times m_1}, \mathbf{b} \in \mathbb{R}^{m_2 \times 1}$

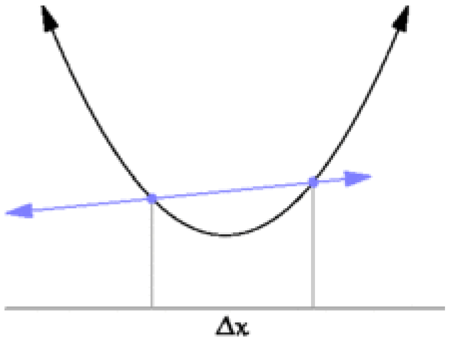
$\mathbf{A} \in \mathbb{R}^{m_2 \times m_1}, \mathbf{B} \in \mathbb{R}^{m_2 \times n}$

Analysis
Differentiation

- Derivative of a function is the slope of the tangent line to the graph of the function.



$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$$



Analysis
Differentiation

- First derivative of a function
  - of a scalar *x*: $$f' = \frac{\mathrm{d}f}{\mathrm{d}x}$$
  - of a vector **x**: $$\nabla_{\mathbf{x}} f = \left[\frac{\partial f}{\partial x_1} \cdots \frac{\partial f}{\partial x_m} \right]^{\mathrm{T}}$$

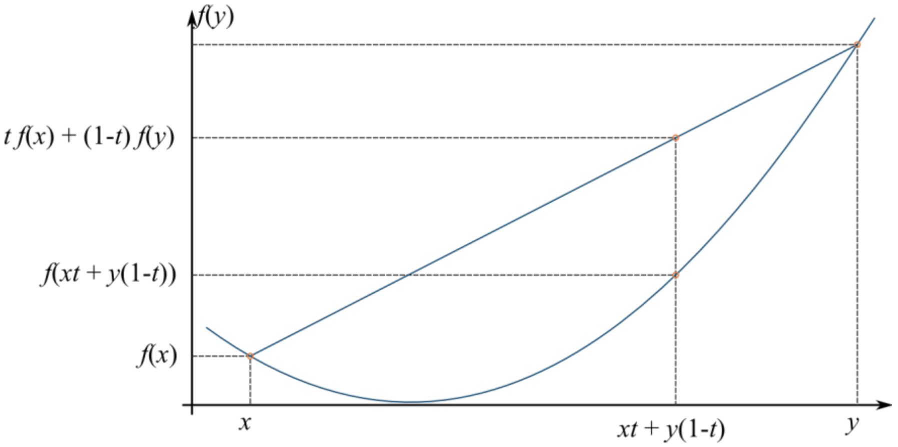
Gradient

Partial derivative

- Second derivative of a function
  - of a scalar *x*: $$f'' = \frac{\mathrm{d}^2 f}{\mathrm{d}x^2}$$
  - of a vector **x**: $$\nabla_{\mathbf{x}}^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_m \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_m^2} \end{bmatrix}$$

Hessian Matrix

Analysis
Convex & concave functions



- Convex function:

$$f(tx+(1-t)y) \le tf(x)+(1-t)f(y)$$

- Concave function:

$$f(tx+(1-t)y) \ge tf(x)+(1-t)f(y)$$

- Strictly convex and concave, resp.:
  - „≤“ and „≥“ become „<“ and „>“.
  - There exist no more than one minimum or maximum, resp.
- Second gradient is non-negative everywhere (non-positive for strictly concave functions)
- Any tangent of *f*(*x*) is a lower bound on f (upper bound for concave functions)

Optimization
Definitions

- Optimization problem (OP):

$$f^* = \min_{x \in S} f(x) \quad \text{with} \quad x^* = \arg\min_{x \in S} f(x)$$

  - *f*: target function.
  - S feasible region (defined by constraints).
  - f^* optimal value.
  - x^* optimal solution.
  - Any $x \in S$ is called *feasible solution*.
- Convex optimization problem:
  - Target function and feasible region are convex.
  - Local Optimum = global Optimum.

Stochastics
Application 1: Diagnostics

- New test has been developed.
- Question: What is the likelihood of a person being sick if the test is positive?
- Study: Apply test on both healthy and sick probands (real state is known).

Stochastics
Application 2: Vaccine

- New vaccine has been developed.
- Question: How good is it? How often does it prevent an infection?
- Study: Test persons are vaccinated and later tested if they got an infection.

What are we investigating?

- *Descriptive statistics*: Describing and investigating attributes of samples.
  - What is the fraction of probands that got an infection? (= counting)
- *Inductive statistics*: Which conclusions regarding the population can be drawn from a sample? (Machine Learning).
  - How many persons will stay healthy in the future?
  - How confident are we regarding that number?

Probabilities

- Frequentist „objective“ probabilities
  - Probabilities as relative frequency of an event in large number of independent and repeated experiments.
- Bayesian „subjective“ probabilities
  - Probabilities as personal belief that an event will appear.
  - Uncertainty translates to lack of information.
    - How likely is it that the vaccination works?
    - New information (e.g. new studies) can change these subjective probabilities.

Probability theory

- *Random experiment*: Defined process in which an observation ω is generated (elementary event / outcome).
- Sample space *Ω*: Set of all possible elementary events. Number of events is |Ω|.
- *Event A*: Subset of sample space .
- *Probability P*: Function that distributes probability mass to events *A* in Ω.

$$P(A) := P\left(\{\omega \in A\}\right)$$

Probability theory

- Probability = *normed measure*
- Defined via Kolmogorov axioms:
  - Probability of event $A \subseteq \Omega$: $0 \leq P(A) \leq 1$
  - Unit measure: $P(\Omega) = 1$
  - Probability of event $A \subseteq \Omega$ or event $B \subseteq \Omega$ with $A \cap B = \varnothing$ (Events are mutually exclusive):
    $$P(A \cup B) = P(A) + P(B)$$
  - In general: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Scheffer/Dick: Language Technology

Random variables

- *Random variable X* is a measurable function from elementary events
  - to numerical value $X : \omega \in \Omega \mapsto x \in \mathbb{R}$
  - or to *m*-dimensional vector $X : \omega \in \Omega \mapsto \mathbf{x} \in \mathbb{R}^m$
  - Machine Learning: Mappings to trees and other structures are also possible.
  - Machine Learning: Used synonymously to sample space.

- Image (or range) of random variable:

$$\mathcal{X} := \{X(\omega) \mid \omega \in \Omega\}$$

Discrete random variable

- X is called a discrete random variable if its set of possible outcomes is discrete.

- Probability function *P* assigns a probability to every possible value of the random variable.

$$P(X = x) \in [0;1]$$

  - Sum of probability function over all values:

$$\sum_{x \in \mathcal{X}} P(X = x) = 1$$

Continuous random variable

- X is a continuous random variable if its set of possible outcomes is continuous.
- The values of the distribution function *P* are defined as the cumulated probabilities

$$P_X(x) = P(X \leq x) \in [0;1]$$

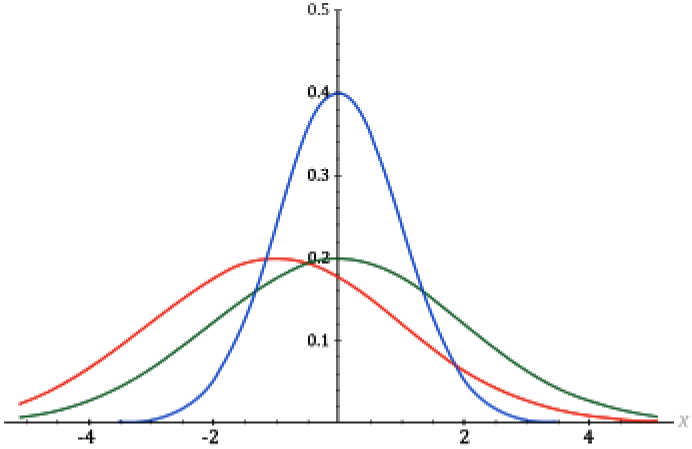
- The values of the probability density function *p* correspond to the change in the distribution function.

$$p_X(a) = \left.\frac{\partial P_X(x)}{\partial x}\right|_{x=a} \qquad \text{with} \qquad \int_{-\infty}^{\infty} p_X(x)\,dx = 1$$

Random variables

- Discrete:
  - E.g. coin toss.

- Continuous:
  - E.g. Gaussian normal distribution.



Notational subtleties

- $P(X)$ p_X — Probability function or probability density function over all values of X

- $P(X = x)$ $p_X(x)$ — specific probability value or specific value of probability density function
- $P(x)$ $p(x)$ — shortened notation of $P(X = x)$ or $p_X(x)$ if the identity of the random variable is unambiguous.

Expectation and variance

- The expected value *E(X)* is the weighted average over all possible values of X
  - Discrete random variable:
  $$E(X) = \sum_{x \in \mathcal{X}} xP(X = x)$$
  - Continuous random variable:
  $$E(X) = \int_{\mathcal{X}} x p_X(x)\, dx$$
- The variance *Var(X)* is the expected quadratic distance to the expected value of *X*

$$Var(X) = E\left[\left(X - E(X)\right)^2\right]$$

Expectation: Example

- St. Petersburg Lottery:
  - Toss a coin until head appears for the first time.
  - Pot starts at 1€.
  - Each time tail appears, the pot is doubled.
  - Value of pot is random variable X.
  - Expected (average) profit:

$$E(X) = \sum_{x \in \mathcal{X}} xP(X = x)$$

$$= 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 4 \cdot \frac{1}{8} + \ldots = \infty$$

  - How much are you willing to pay to enter the game?

Joint Probability

- $P(X_1, X_2)$ is the joint probability distribution of random variables X_1 and X_2
- Joint image:

Cartesian product

- E.g.:

$$\mathcal{X}_1 \times \mathcal{X}_2$$

$\mathcal{X}_1 \times \mathcal{X}_2 =$ { (sick, sick), (sick, healthy), (healthy, sick), (healthy, healthy) }

Conditional Probabilities

- Conditional Probability: Probability of values of X with additional information:
  - Discrete random variable:

$$P\left(X = x \mid \text{Additional Information}\right)$$

  - Continuous random variable:

$$p_X\left(x \mid \text{Additional Information}\right)$$

- Definition of conditional probability:

$$P(X|Y = y) = \frac{P(X,Y=y)}{P(Y=y)}$$

Rules for Calculating Probabilities

- Product rule:

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

  - General product rule (chain rule):

    $$P(X_1, X_2, \ldots, X_n) = P(X_1) \prod_{i=2}^{n} P(X_i | X_1, \ldots, X_{i-1})$$

- Sum rule:
  - If two events, A and B, are mutually exclusive:

    $$P(A \cup B) = P(A) + P(B)$$

- Marginal distribution:

$$P(X) = \sum_{y \in \mathcal{Y}} P(X, Y = y) = \sum_{y \in \mathcal{Y}} P(X|Y = y)P(Y = y)$$

Rules for Calculating Probabilities

- Bayes' theorem:
  - Infer $P(X \mid Y)$ from $P(Y \mid X)$, $P(X)$, and $P(Y)$

$$P(X,Y) = P(Y,X)$$

$$\Leftrightarrow P(X \mid Y)P(Y) = P(Y \mid X)P(X)$$

$$\Leftrightarrow P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

Dependent Random Variables

- Random variables X_1 and X_2 can either be dependent or independent.
- Independent: $P(X_1, X_2) = P(X_1)\, P(X_2)$
  - Example:
    - 2 consecutive coin tosses (fair coin).
    - Result of second event does not dependent on first event.
  - Implies: $P(X_2 \mid X_1) = P(X_2)$
- Dependent: $P(X_1, X_2)= P(X_1)\, P(X_2 \mid X_1) \neq P(X_1)\, P(X_2)$
  - Example:
    - Flu symptoms of 2 people sitting next to each other.

Conditional Independence

- Random variables can be dependent and at the same time independent given another random variable.
- The random variables X_1 and X_2 are conditional independent given Y if:
  - $P(X_1,X_2|Y) = P(X_1|Y)\ P(X_2|Y)$
- Example:
  - Effectiveness of vaccinate known → probabilities of infections independent
  - Effectiveness of vaccinate unknown → Observation of probands gives clues for other probands.

Application 1: Diagnostics

- New test has been developed.
- Question: What is the likelihood of a person being sick if the test is positive?
- Study: Apply test on both healthy and sick probands (real state is known).

Application 2: Vaccine

- New vaccine has been developed.
- Question: How good is it? How often does it prevent an infection?
- Study: Test persons are vaccinated and later tested if they got an infection.

Bayes' Theorem: Example

- Diagnostics example:
  - $P(positive \mid sick) = 0.98$
  - $P(positive \mid healthy) = 0.05$
  - $P(sick) = 0.02$
- Given test result *Test*, we want to know:
  - Probability that the patient is sick:

    $$P(sick \mid Test)$$

  - Most plausible cause

    $$\arg\max_{S \in \{sick, healthy\}} P(Test \mid S)$$

  - Most probable cause

    $$\arg\max_{S \in \{sick, healthy\}} P(S \mid Test)$$

Bayes‘ Theorem

- Probability of real cause *Cau.* for observation *Obs.*:

$$P(\text{Cau} \mid \text{Obs}) = P(\text{Obs} \mid \text{Cau}) \frac{P(\text{Cau})}{P(\text{Obs})}$$

$$P(\text{Obs}) = \sum_{c \in Causes} P(\text{Obs} \mid c) P(c)$$

- $P(\text{Cau})$: Prior probability, „Prior“.
- $P(\text{Obs}|\text{Cau})$: Likelihood.
- $P(\text{Cau}|\text{Obs})$: Poster probability, „Posterior“.

Prior, Likelihood, and Posterior

- Subjective estimate, **before** we have seen any data: prior distribution over models
  - $P(Health)$
  - $P(\theta)$, θ – effectiveness of vaccination
- How well does data fit to model: Likelihood
  - $P(Test \mid Health)$
  - $P(Study \mid \theta)$,
- Subjective estimate, **after** we have seen data: posterior distribution
  - $P(Health \mid Test)$
  - $P(\theta \mid Study)$

Prior

- Where do we get a prior distribution from?
  - $P(Health)$ relatively easy; discrete.
  - $P(\theta)$: harder; continuous; could e.g. be estimated from all current studies on other vaccinations.
- By definition, a prior expresses one‘s belief about a random variable. There is no ‚correct‘ prior.
  - But: Choice of prior distribution influences the quality of future predictions.
- Posterior distribution is computable from prior and likelihood of the observations.
  - using Bayes‘ theorem

Example for Likelihood: Bernoulli Distribution

- A discrete distribution with two possible outcomes 0 and 1 is a Bernoulli distribution.
- Determined by exactly one parameter:

$$\theta \in [0; 1]$$

- Distribution function:

$$P(X = 1|\theta) = \theta$$
$$P(X = 0|\theta) = 1 - \theta$$

Example for Likelihood: Binomial Distribution

- Collection of several Bernoulli distributed random variables $X_1, \ldots, X_n$ with same parameter θ.
  - New random variable Y, which determines how many of the X_i are positive: $$Y = \sum_{i=1}^{n} X_i$$
  - Y is binomially distributed with parameters θ and n
  - Distribution function:

$$P(Y = y \mid \theta, n) = \binom{n}{y} \theta^{y} (1-\theta)^{n-y}$$

Binomial coefficient: number of possibilities to draw y elements out of a set of n elements.

Probability that y random variables X_i are positive.

Probability that n-y random variables X_i are negative.

Example for Prior: Beta Distribution

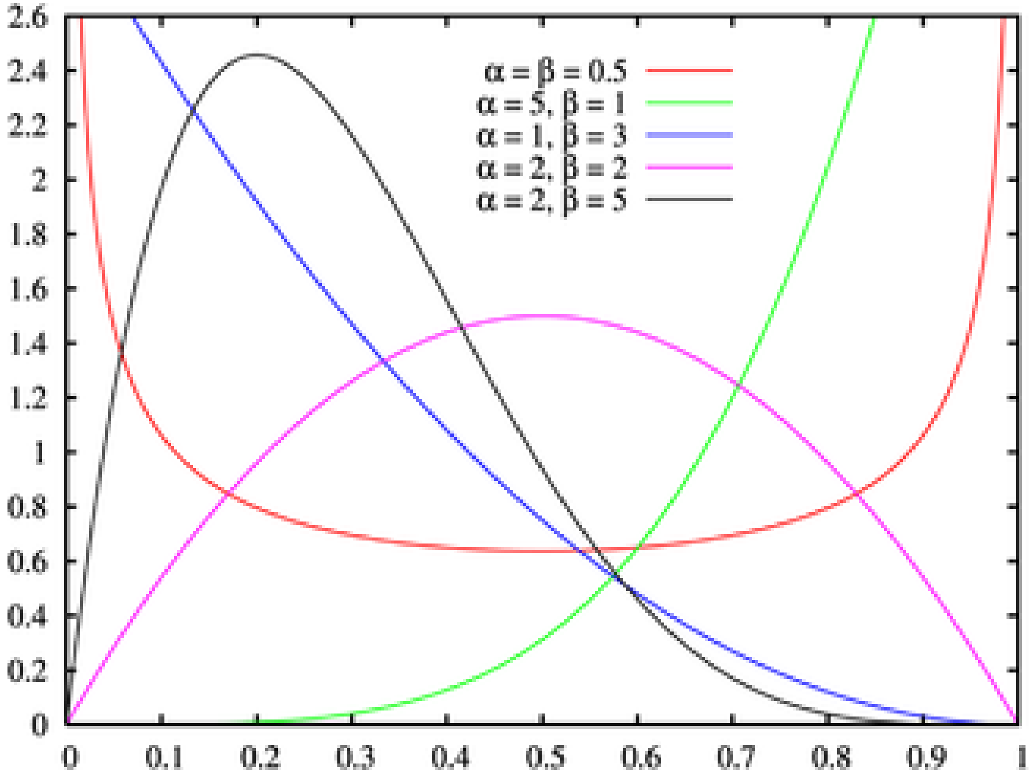
- Distribution over all possible effectiveness rates.
- Continuous distribution.
- $P(\theta)$ is a density function

- Common choice (with parameter $\theta \in [0; 1]$):
  - Beta distribution
  - defined by 2 parameters α and β

$$P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

Beta function; used for normalization

Example for Prior: Beta Distribution



- Special case: $\alpha = \beta = 1$ is uniform distribution

$$P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} = \frac{\theta^0(1-\theta)^0}{1} = 1$$

General Pattern for Computation of the Posterior Distribution

- We have:
  - Prior distribution $P(\theta)$
  - Observation $x_1, \ldots, x_n$
  - Likelihood $P(x_1, \ldots, x_n \mid \theta)$
- We want: Posterior distribution $P(\theta \mid x_1, \ldots, x_n)$
- 1. Apply Bayes‘ theorem.
  $$P(\theta|x_1, \ldots, x_n) = P(x_1, \ldots, x_n|\theta)P(\theta)/P(x_1, \ldots, x_n)$$
- 2. Apply marginal distribution for continuous parameters.
  $$P(x_1, \ldots, x_n) = \int P(x_1, \ldots, x_n|\theta)P(\theta)d\theta$$

Computation of the Posterior Distribution: Practical Example

- Given:
  - Model parameter space $\theta \in [0;1]$
  - Beta prior with parameters α and β:
    $P(\theta)=Beta(\theta \mid \alpha,\beta)$
  - Bernoulli likelihood
  - Binary observations $x_1, \ldots, x_n$, conditionally independent given model parameter θ
    - ★ a positive observations, b negative
- Compute:
  - Posterior $P(\theta \mid x_1, \ldots, x_n)$

Computation of the Posterior Distribution

$$P(\theta \mid x_1,\ldots,x_n)$$

$$= P(x_1,\ldots,x_n \mid \theta) P(\theta) / P(x_1,\ldots,x_n)$$ Bayes' theorem

$$= \left[\prod_{i=1}^{n} P(x_i \mid \theta)\right] P(\theta) / P(x_1,\ldots,x_n)$$ Conditional independence

$$= P(X=1 \mid \theta)^a P(X=0 \mid \theta)^b P(\theta) / P(x_1,\ldots,x_n)$$ a positive, b negative

$$= \theta^a (1-\theta)^b \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} / P(x_1,\ldots,x_n)$$ Bernoulli and Beta distributions

$$= \frac{\theta^{a+\alpha-1}(1-\theta)^{b+\beta-1}}{B(\alpha,\beta)} / \left[\int \frac{\theta^{a+\alpha-1}(1-\theta)^{b+\beta-1}}{B(\alpha,\beta)} d\theta\right]$$ Shorten expressions, marginal distribution formula

$$= \frac{\theta^{a+\alpha-1}(1-\theta)^{b+\beta-1}}{B(\alpha,\beta)} / \left[\frac{B(a+\alpha, b+\beta)}{B(\alpha,\beta)}\right]$$ Definition of Beta function

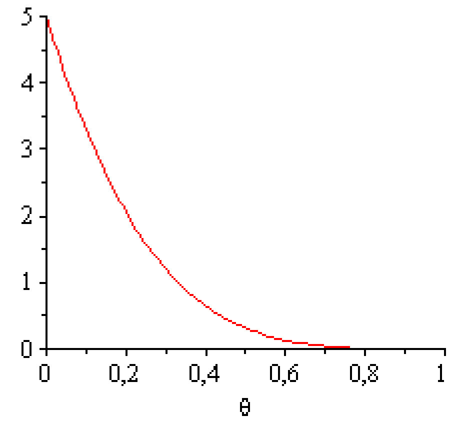
$$= Beta(\theta \mid a+\alpha, b+\beta)$$ Canceling, Definition of Beta distribution

Conjugate Prior

- Previous example:
  - Starting from prior $Beta(\theta \mid \alpha, \beta)$
  - using *a* positive and *b* negative observations
  - we computed posterior $Beta(\theta \mid \alpha + a, \beta + b)$
  - Algebraic forms of posterior and prior are identical.
- Beta distribution is conjugate prior of Bernoulli likelihood.
- It is generally good to use the conjugate prior, in order to guarantee that the posterior is efficiently computable.

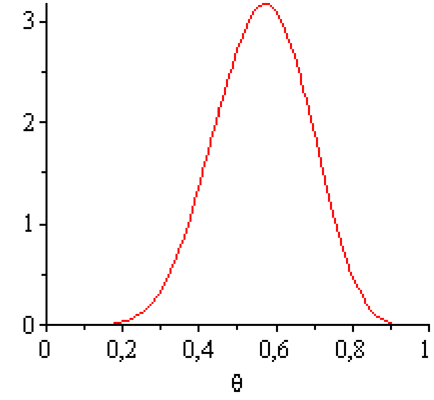
Practical Example: Vaccination Study

- Prior: Beta with α=1, β=5



- 8 healthy probands, 2 infected
- Corresponding posterior: Beta with α=9, β=7

- Parameters of Beta distribution take role of pseudo counts.



Prediction / Inference

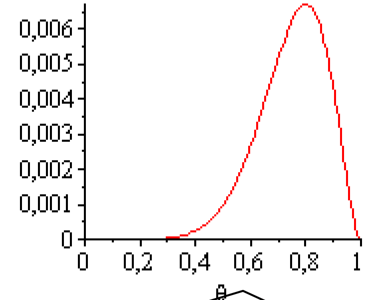
- Which observations can we expect in the future, given our belief about the probability distribution?
  - Prediction of test data, given distribution parameters, e.g. $P(X_{new} | \hat{\theta})$, e.g. belief that vaccination effectiveness is $\hat{\theta} = 0.7$
  - or $P(X_{new}) = \int_{\theta} P(X_{new} | \theta) P(\theta)$, e.g. belief that vaccination effectiveness is Beta distributed with (9,7)
- Which observations can we expect in the future, given past observation?
  - Prediction of test data, given a set of training data $P(X_{new} | X_{old})$. This is also called inference in graphical models (next lecture).

Parameter Estimation

- Bayesian inference doesn‘t yield model parameters but distribution over model parameters.
- Estimation of model with highest probability: MAP estimation
  - „maximum-a-posteriori“ = maximizes the posterior
  - $\theta_{MAP} = \text{argmax}_{\theta} P(\theta \mid observations)$
- In contrast: most *plausible* model = ML estimation
  - „maximum-likelihood“ = maximizes likelihood
  - without considering Priors
  - $\theta_{ML} = \text{argmax}_{\theta} P(observations \mid \theta)$

Parameter Estimation: Example

- Vaccination study:
  - Prior: Beta with α=1, β=5
  - 8 healthy probands, 2 infected
  - Corresponding posterior: Beta with α=9, β=7
- ML estimation:
  - $\theta_{ML} = \text{argmax}_\theta P(Obs \mid \theta)$
  - $$\theta_{ML} = \arg\max_\theta \theta^8(1-\theta)^2 = \frac{4}{5}$$



Likelihood function
(no probability distribution)

- MAP estimation:
  - $\theta_{MAP} = \text{argmax}_\theta P(\theta \mid Obs)$

$$\theta_{MAP} = \arg\max_\theta \frac{\theta^8(1-\theta)^6}{B(9,7)} = \frac{4}{7}$$

Parameter Estimation: MAP

- We want: The parameter that maximizes the posterior distribution $P(\theta \mid x_1, \ldots, x_n)$.
- Before: Compute posterior distribution.
  - 1. Apply Bayes‘ theorem.

$$P(\theta | x_1, \ldots, x_n) = P(x_1, \ldots, x_n | \theta) P(\theta) / P(x_1, \ldots, x_n)$$

  - 2. Apply marginal distribution for continuous parameters.

$$P(x_1, \ldots, x_n) = \int P(x_1, \ldots, x_n | \theta) P(\theta) d\theta$$

- We don‘t need the marginal distribution $P(x_1, \ldots, x_n)$ to compute the MAP parameter!

Prediction / Inference

- Which observations can we expect in the future, given past observation?
  - Prediction of test data, given a set of training data $P(X_{new} \mid X_{old})$
- Prediction using MAP estimation:
  - Compute θ_{MAP} via $\theta_{MAP} = argmax_{\theta} P(\theta \mid X_{old})$
  - Then compute $P(X_{new} \mid \theta_{MAP})$ (Likelihood distribution)
  - Loss of information:
    - θ_{MAP} is not the „real“ parameter but the most likely.
    - Approach ignores that other models are also possible.

Bayes Optimal Prediction

- No intermediate step using the MAP model. Instead, direct derivation of the prediction:

$$P(X_{new} \mid X_{old})$$

1. Marginal distribution

$$= \int_{\theta} P(X_{new} \mid \theta, X_{old}) P(\theta \mid X_{old}) d\theta$$

2. Conditional independence

$$= \int_{\theta} P(X_{new} \mid \theta) P(\theta \mid X_{old}) d\theta$$

Average over *all* models (Bayesian Model Averaging)

Prediction given model

Weighted by how good model fits to previous observations. (Posterior)

Prediction: Example

- Vaccination study: What is the probability of a person staying healthy, given the study?
- Prediction using MAP model:
  - $\theta_{MAP} = \text{argmax}_{\theta} P(\theta \mid Obs) = 4/7$
  - $P(healthy \mid \theta_{MAP}) = \theta_{MAP} = 4/7$
- Bayes optimal prediction:

$$P\left(healthy \mid X_{old}\right) = \int_{\theta} P\left(healthy \mid \theta\right) P\left(\theta \mid X_{old}\right) d\theta$$

$$= \int_{\theta} \theta \cdot Beta\left(\theta \mid 9,7\right) d\theta = \frac{9}{16}$$

Expected value of Beta distribution

Summary

- Bayesian Learning:
  - Prior: subjective start distribution over models
  - Past observations: Likelihood given model parameters
  - With Bayes‘ theorem: Posterior: Distribution over models given data.
  - Possible ways of future predictions:
    - Compute MAP model (maximization of posterior), afterwards prediction with MAP Model ← simpler
    - Bayes optimal prediction: average over all models, weighted with posterior. ← better

Questions?