Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Bayesian Learning

Niels Landwehr

Overview

- Basic concepts of Bayesian learning
- Most probable model given data
  - Coin tosses
  - Linear regression
  - Logistic regression
- Bayesian predictions
  - Coin tosses
  - Linear regression

Overview

- Basic concepts of Bayesian learning
- Most probable model given data
  - Coin tosses
  - Linear regression
  - Logistic regression
- Bayesian predictions
  - Coin tosses
  - Linear regression

Statistics & Machine Learning

- Machine Learning: tightly related to (inductive) Statistics
- Two areas in Statistics:
  - *Descriptive Statistics*: description and examination of the properties of data.

Mean values

Variances

Difference between Populations

  - *Inductive Statistics*: What conclusions can be drawn from data about the underlying reality?

Model building

Explanations for observations

Relationships and patterns in the data

Thomas Bayes

- 1702-1761
- “An essay towards solving a problem in the doctrine of chances“, published in 1764.

- The work of Bayes laid the foundations for inductive Statistics.

- “Bayesian Probabilities“ offer an important perspective on uncertainty and probability.

Frequentist / Bayesian Probabilities

- Frequentist probabilities
  - Describe the possibility of an occurrence of an intrinsically stochastic event (e.g. a coin toss).
  - Defined as *relative frequencies* of possible outcomes in a *repeatable experiment*

"If one throws a fair coin 1000 times,
it will land on heads about 500 times"

"In 1 gram of Potassium-40, around 260,000 nuclei decay per second"

Frequentist / Bayesian Probabilities

- Bayesian “subjective“ probabilities
  - Here, the reason for uncertainty is attributed to a lack of information.
  - How likely is it that suspect X killed the victim?
  - New Information (e.g., finger prints) can change these subjective probabilities.

- Bayesian view is more important in machine learning
- Frequentist perspective is also sometimes used, or is mathematically equivalent to the Bayesian approach

Bayesian Probability in Machine Learning

- Model building: find an explanation for observations
- What is the “most probable“ model? Trade-off between
  - Prior knowledge (a priori distribution over models)
  - Evidence (data, observations)
- Bayesian Perspective:
  - Evidence (data) changes the “subjective“ probability for models (explanation)
  - A posteriori model probability, MAP hypothesis

Conceptual Model for Learning

- Many machine learning methods are based on probabilistic considerations.
- We want to learn models of the form $y = f(\mathbf{x})$ from training data $L = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.
- conceptual model of the data generating process
  - Someone draws the real model f^* from the ("prior") distribution $p(f)$.
  - f^* is not known, but $p(f)$ reflects prior knowledge (what are the most probable models?)
  - Training inputs $\mathbf{x}_i$ are drawn (independent of $\boldsymbol{\theta}^*$).
  - Class labels y_i are drawn from $p(y_i|\mathbf{x}_i, \boldsymbol{\theta}^*)$.
  - **Learning Question**: given L and $p(\boldsymbol{\theta})$, what is the most likely "true" model?
    → Try to (approximately) reconstruct $\boldsymbol{\theta}^*$

Bayes‘ Rule

- Bayes‘ Rule:

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

- Proof is simple:

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

Definition of conditional distribution

Product rule

- Important basic knowledge for machine learning: allows the inference of *model probabilities* given the *probabilities of observations*

Bayes‘ Rule

- Model probability given data and prior knowledge

Likelihood: how probable is the data, under the assumption that “model“ is the true model?

Prior: how probable is a model, a priori?

$$p(model|data) = \frac{p(data|model)p(model)}{p(data)}$$

$p(data)$ is constant; it is independent of model

$$\propto p(data|model)p(model)$$

Most probable model given data

- Given
  - assumed a priori distribution $p(f)$
  - observed data $L = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$
- Question: what is the most probablye model, given the assumed a priori distribution and the data?
- So-called MAP („maximum a posteriori“) model.

$$f^* = \arg\max_f \; p(f \mid L)$$

$$= \arg\max_f \; \frac{p(L \mid f)\, p(f)}{p(L)} \quad \text{Bayes rule}$$

$$= \arg\max_f \; p(L \mid f)\, p(f) \quad p(L) \text{ independent of } f$$

Most probable model given data

- To compute the MAP model, we need two terms:
  - Likelihood $p(L \mid f)$: Probability of the data given the model. Follows from definition of model.
  - Prior $p(f)$: A-priori distribution over models. Assumption, usually reflects prior knowledge.
    Example learning of decision trees: „Small trees are often better than large trees“.
- We will go through several examples later on.

Overview

- Basic concepts of Bayesian learning
- Most probable model given data
  - Coin tosses
  - Linear regression
  - Logistic regression
- Bayesian predictions
  - Coin tosses
  - Linear regression

Estimating a model for coin tosses

- Application of Bayes‘ rule on a simple example: estimate model for coin tosses („binomially distributed data“).

- Coin toss model
  - Tossing a coin N times.
  - Result of i-th coin toss: „Head“ or “Tails“, random variable $y_i \in \{0,1\}$.
  - y_i follow a Bernoulli distribution:

$$p(y_i = 1) = \theta \qquad \text{Probability for „Head“}$$

$$p(y_i = 0) = 1 - \theta \qquad \text{Probability for „Tails“}$$

$$y_i \sim \text{Bern}(y_i \mid \theta) = \theta^{y_i}(1-\theta)^{1-y_i} \qquad \text{Bernoulli distribution}$$

- Data are also called „binomially distributed“, because the number of „Heads“in N coin tosses follows binomial distribution.

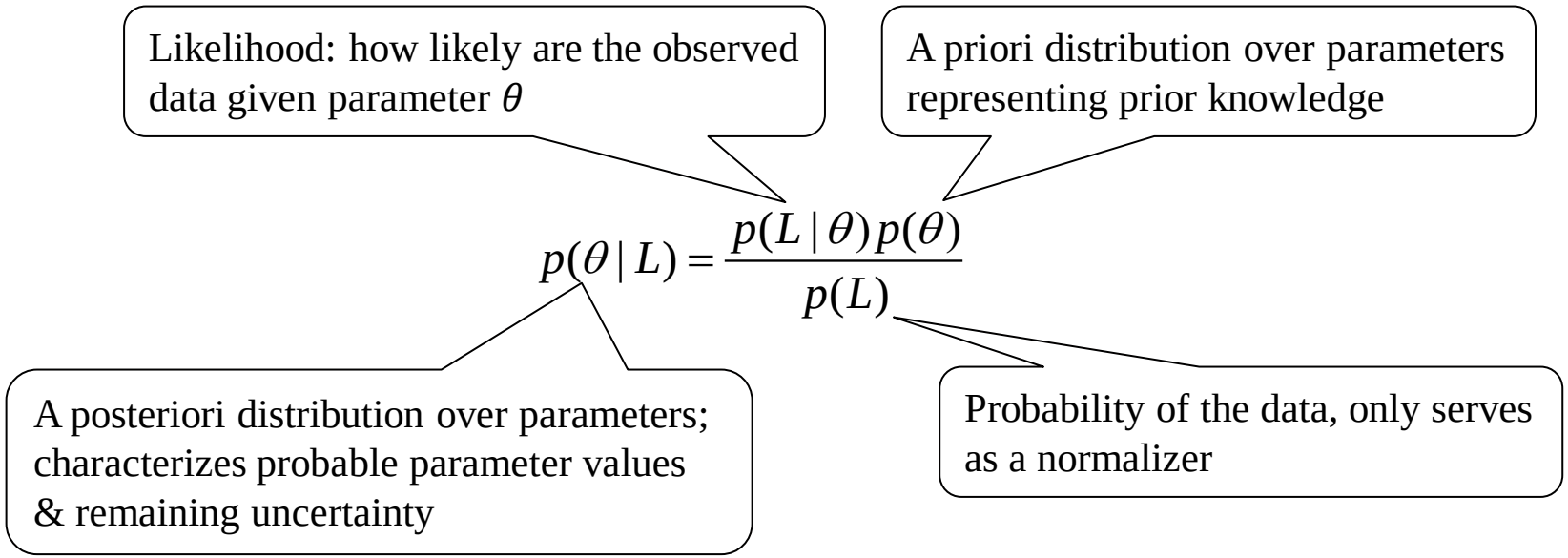
Estimating a model for coin tosses

- The coin toss model predicts the probability of seeing „Head“ in a single coin toss.
- Model is specified by a single parameter: Probability $\theta \in [0,1]$ for observing a „Head“.
- In contrast to our standard learning problem setting there are no feature representations $\mathbf{x}_i$ of instances.

- We would like to estimate a coin toss model from observations $L = \{y_1, ..., y_N\}$.
- The „true “ probability θ for seeing a „Head“ we cannot know.
- However, with Bayes‘ rule we can derive probabilistic statements about this true probability, given data.

Bayes‘ rule for coin tosses

- Derivation with Bayes‘ rule:



Likelihood coin tosses

- Likelihood: how likely are the data $L = \{y_1, ..., y_N\}$ given a model θ ?
- Define $N_k = \sum_{i=1}^{N} y_i$ (number of „Heads“), $N_z = N - N_k$ (number of „Tails“).
- Likelihood is given by

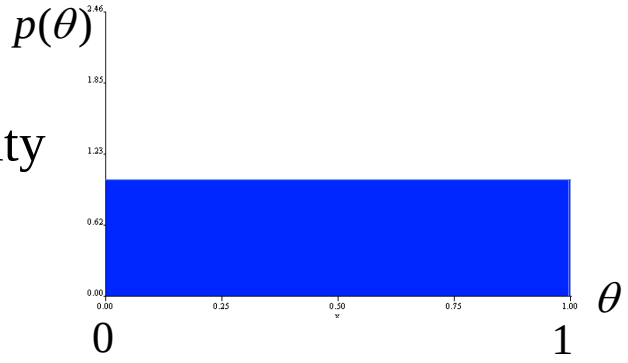
$$p(L \mid \theta) = \prod_{i=1}^{N} p(y_i \mid \theta) = \theta^{N_k} (1-\theta)^{N_z}.$$

coin tosses independent

Prior coin tosses: uniform

- How to define the prior $p(\theta)$ for coin tosses?
- 1) Try: no prior knowledge

$$p(\theta) = \begin{cases} 1 : 0 \le \theta \le 1 \\ 0 : sonst \end{cases}$$



- Example:
  - Data L = {Tails, Tails, Tails}
  - MAP model:

$$\theta^* = \arg\max_{\theta \in [0,1]} p(\theta \mid L) = \arg\max_{\theta \in [0,1]} \frac{p(L \mid \theta)\, p(\theta)}{p(L)}$$

$$= \arg\max_{\theta \in [0,1]} p(L \mid \theta) = \arg\max_{\theta \in [0,1]} \theta^0 (1-\theta)^3 = 0$$

- Conclusion: coin will **never** show „Heads“
  - Bad, „overfitting“ to data

Prior coin tosses: Beta

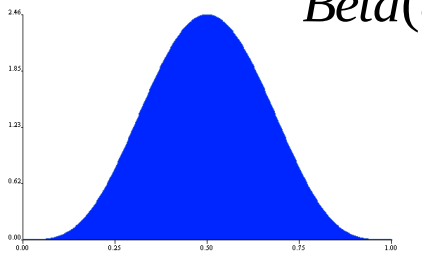
- How to define the prior $p(\theta)$ for coin tosses?
- Better use prior knowledge: Unlikely that coin will always show „Heads“ or always „Tails“.
- Good model for prior knowledge: Beta distribution.

$$p(\theta) = Beta(\theta \mid \alpha_k, \alpha_z)$$

$$= \frac{\Gamma(\alpha_k + \alpha_z)}{\Gamma(\alpha_k)\Gamma(\alpha_z)} \theta^{\alpha_k - 1}(1-\theta)^{\alpha_z - 1}$$

$$(\theta \in [0,1])$$

$Beta(\theta \mid 5,5)$



0.5

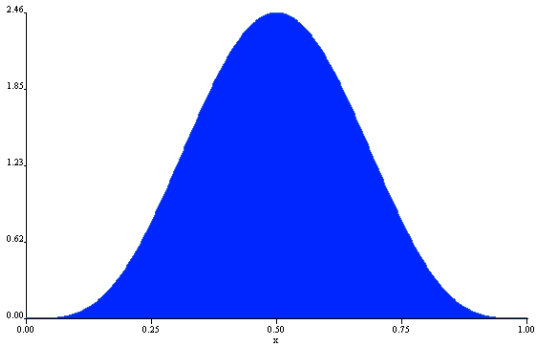
- Gamma function $\Gamma(\alpha)$ is the continuous extension of the factorial function.

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt \qquad \forall n \in \mathbb{N} : \Gamma(n) = (n-1)!$$

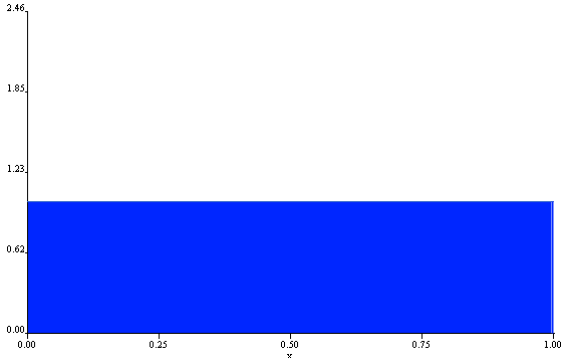
Beta distribution

- α_k and α_z are parameters of the Beta distribution („hyperparameters“)

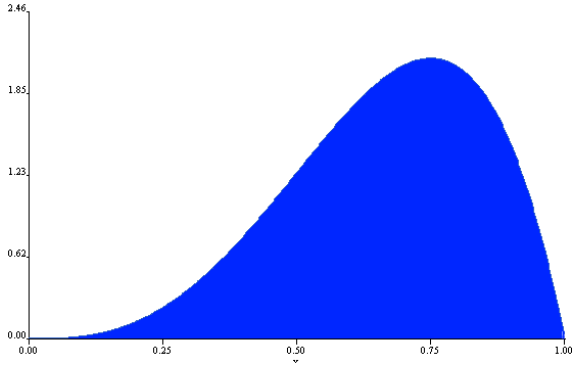
$\alpha_K = 5, \quad \alpha_Z = 5$



$\alpha_K = 1, \quad \alpha_Z = 1$



$\alpha_K = 4, \quad \alpha_Z = 2$



- Normalized density:

$$\int_0^1 Beta(\theta \mid \alpha_K, \alpha_Z) d\theta = 1.$$

Conjugate prior

- Why choose this particular prior distribution?
- Structural similarity with likelihood:

Prior $$p(\theta) = Beta(\theta \mid \alpha_k, \alpha_z) = \frac{\Gamma(\alpha_k + \alpha_z)}{\Gamma(\alpha_k)\Gamma(\alpha_z)} \theta^{\alpha_k - 1}(1-\theta)^{\alpha_z - 1}$$

Likelihood $$p(L \mid \theta) = \theta^{N_k}(1-\theta)^{N_z}$$

- Easy to incorporate observations: product of likelihood and prior has the same form as prior

$$p(\theta \mid L) \propto p(L \mid \theta)\, p(\theta)$$

Conjugate prior

- If we plug the beta prior into Bayes‘ equation, we obtain:

$$p(\theta \mid L) = \frac{p(L \mid \theta)\, p(\theta)}{p(L)}$$

$$= \frac{1}{Z} \theta^{N_k} (1-\theta)^{N_z} \frac{\Gamma(\alpha_k + \alpha_z)}{\Gamma(\alpha_k)\Gamma(\alpha_z)} \theta^{\alpha_k - 1} (1-\theta)^{\alpha_z - 1}$$

$$= \frac{1}{Z'} \theta^{\alpha_k + N_k - 1} (1-\theta)^{\alpha_z + N_z - 1}$$

$$= ?$$

What is the normalizer Z‘?

Conjugate prior

- If we plug the beta prior into Bayes‘ equation, we obtain:

$$
\begin{aligned}
p(\theta \mid L) &= \frac{p(L \mid \theta)\, p(\theta)}{p(L)} \\
&= \frac{1}{Z}\theta^{N_k}(1-\theta)^{N_z}\,\frac{\Gamma(\alpha_k+\alpha_z)}{\Gamma(\alpha_k)\Gamma(\alpha_z)}\theta^{\alpha_k-1}(1-\theta)^{\alpha_z-1} \\
&= \frac{1}{Z'}\theta^{\alpha_k+N_k-1}(1-\theta)^{\alpha_z+N_z-1} \\
&= \frac{\Gamma(\alpha_k+N_k+\alpha_z+N_z)}{\Gamma(\alpha_k+N_k)\Gamma(\alpha_z+N_z)}\theta^{\alpha_k+N_k-1}(1-\theta)^{\alpha_z+N_z-1} \\
&= Beta(\theta \mid \alpha_k+N_k, \alpha_z+N_z)
\end{aligned}
$$

- **Beta distribution is „conjugate“ Prior: Posterior again follows a Beta distribution.**

Summary computation of posterior

- Summary computation of posteriori distribution:
- Derive with: Bayes‘ rule

$$p(\theta \mid L) = \frac{p(L \mid \theta)\, p(\theta)}{p(L)}$$

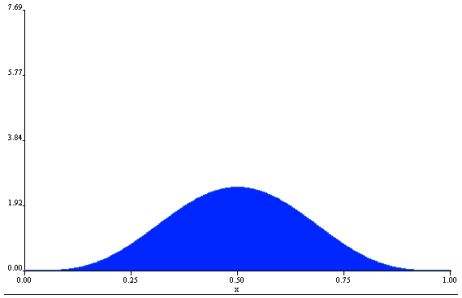
- Determine likelihood, choose appropriate prior distribution (Beta).
- Posterior $p(\theta \mid L)$: How probable is model θ after having seen the data L? That is, how probable is it that θ is the „true“ probability for „Head“?

- Computation of posterior easy in this case:
  - Prior Beta($\theta \mid \alpha_k, \alpha_z$).
  - Observing N_k „Heads“ and N_z „Tails“.
  - Resulting posterior Beta($\theta \mid \alpha_k + N_k, \alpha_z + N_z$).

Coin tosses: example

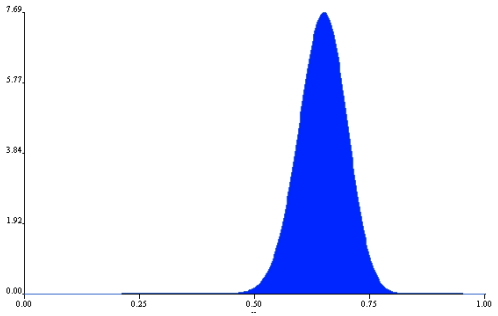
- Example computation of posterior

Prior $\mathrm{p}(\theta) = Beta(\theta \mid 5,5)$



Posterior after L={50x Heads, 25x Tails}:

$$p(\theta \mid L) = Beta(\theta \mid 5+50, 5+25)$$
$$= Beta(\theta \mid 55, 30)$$



$N_k = 50, N_z = 25, \alpha_k = 5, \alpha_z = 5$

Coin tosses: most probable model

- Most probable model given the data:

$$\theta^* = \arg\max_\theta \, p(\theta \mid L)$$

$$= \arg\max_\theta \, Beta(\theta \mid \alpha_k + N_k, \alpha_z + N_z)$$

$$= \arg\max_\theta \, \frac{\Gamma(\alpha_k + \alpha_z + N_k + N_z)}{\Gamma(\alpha_k + N_k)\Gamma(\alpha_z + N_z)} \theta^{\alpha_k + N_k - 1}(1-\theta)^{\alpha_z + N_z - 1}$$

normalizer, independent of θ

$$= \arg\max_\theta \, \theta^{\alpha_k + N_k - 1}(1-\theta)^{\alpha_z + N_z - 1}$$

$$= \frac{N_k + \alpha_k - 1}{N_k + N_z + \alpha_k + \alpha_z - 2}$$

Set derivative to zero
($\alpha_z \geq 1,\ \alpha_k \geq 1$)

- θ^* is the most probable value for the true probability of seeing „Heads“ (given prior, data).

Coin tosses: most probable model

- Most probable model given the data:

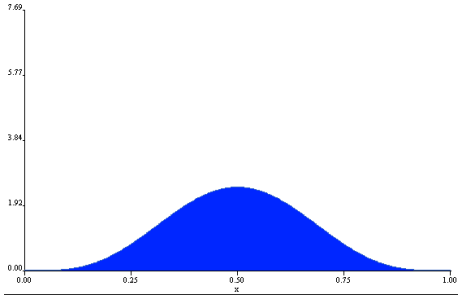
$$\theta^* = \frac{N_k + \alpha_k - 1}{N_k + N_z + \alpha_k + \alpha_z - 2}$$

- Interpretation
  - For $\alpha_z = \alpha_k = 1$ the prior is uniform, and we estimate the maximum likelihood model.
  - Hyperparameters $\alpha_z - 1 / \alpha_k - 1$ of the Beta distribution: „Pseudocounts“ that are added to the observed „counts“ N_z / N_k .
  - How often in our life did we already see a coin toss landing on „Heads“/“Tails“?

Coin tosses: example

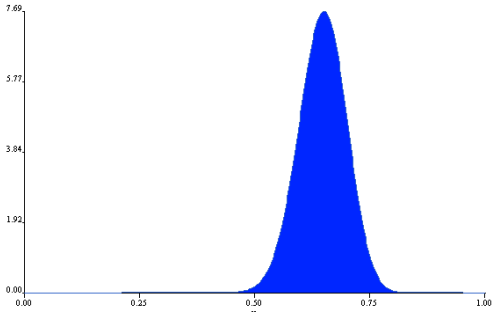
- Example MAP estimate of coin toss parameter.

Prior $\mathrm{p}(\theta) = Beta(\theta \mid 5,5)$



Posterior after L={50x Heads, 25x Tails}:

$$p(\theta \mid L) = Beta(\theta \mid 5+50, 5+25)$$
$$= Beta(\theta \mid 55, 30)$$



$N_k = 50, N_z = 25, \alpha_k = 5, \alpha_z = 5$

MAP estimate: $\theta^* = \arg\max_\theta \; p(\theta \mid L) = \dfrac{54}{54+29} \approx 0.65$

Overview

- Basic concepts of Bayesian learning
- Most probable model given data
  - Coin tosses
  - Linear regression
  - Logistic regression
- Bayesian predictions
  - Coin tosses
  - Linear regression