Bachelor Thesis, Mathematics

“Machine Learning using Bayesian statistics and Neural networks”

A thesis submitted in fulfillment of the requirements for the bachelor’s degree in Mathematics

Author: Lukas Prokop
Supervisor: 高山・信毅 ※
Supervisor: Bredies Kristian †

※ University of Kobe, Japan
† University of Graz, Austria

Version: December 13, 2016

Abstract

Abstract to be done...

Keywords: Machine learning, Bayesian statistics, Bayes' Theorem, Probability theory, Neural networks

Abstract

Abstract in Japanese to be done...

Keywords: Machine learning, Bayesian statistics, Bayes' Theorem, Probability theory, Neural networks

Acknowledgements

Acknowledgements to be done...

どうもありがとうございました。

All source codes are available at lukas-prokop.at/proj/nn and published under terms and conditions of Free/Libre Open Source Software. This document was printed with XeLaTeX in the Andada typeface.

Contents

Chapter 1

Bayesian theory

1.1 Probability theory and statistics

1.1.1 σ-algebra

The following object is necessary in order to define probability properly:

Definition 1 Let X be a set. A σ-algebra is a set Y of subsets of X satisfying:

1. $\emptyset \in Y$
2. $Z \in Y \implies Z^C \in Y$ where Z^C denotes the complement of Z.
3. $(\bigcup_{i=1}^{n} Z_i) \in Y$ where $n \in \mathbb{N}$ and $Z_i \in Y$ where $1 \leq i \leq n$.

Example 1 Let $X \coloneqq \{a, b, c, d\}$ and $Z \coloneqq \{\{a\}\}$. We extend Z to a σ-algebra Y:

$$Y = \{\{\}, \{a, b, c, d\}, \{a\}, \{b, c, d\}\}$$

1.1.2 Basic definitions

Probability theory is concerned with random experiments and random phenomena. Probability in its basic form is the fraction of events with a certain outcome to the total number of events.

Definition 2 A *probability space* $(\Omega, \mathcal{A}, P)$ denotes the set of possible outcomes, a set of events and a map from an element of $\mathcal{A}$ to a real value in $[0, 1]$ respectively. An *event* is an arbitrary set of elements in Ω satisfying the conditions of a σ-algebra. A *random variable* can realize one of the values in $\mathcal{A}$.

Example 2 We toss a coin two times with possible outcomes heads (*h*) or tails (*t*). Then $\Omega = \{h, t\}$ and $\mathcal{A} = \{(h, h), (h, t), (t, h), (t, t)\}$. Let R be our random variable. Our first experiment yields (h, t) as result. Our second experiment yields (h, h). Then we denote $\mathbb{P}[R = (h, t)] \coloneqq 0.5$, $\mathbb{P}[R = (h, h)] \coloneqq 0.5$ and $\mathbb{P}[R = (t, t)] \coloneqq 0$.

1.1.3 Average Value and Expected Value

The *average value* μ of a random variable is defined as,

$$\mu \coloneqq \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} a \tag{1.1}$$

$\mathbb{E}[R]$ (also called *population mean* μ) denotes the expected value of random variable R.

$$\mathbb{E}[R] \coloneqq \sum_{a \in \mathcal{A}} \mathbb{P}[R = a] \cdot a \tag{1.2}$$

So, average and expected values are only defined for numeric events, unlike our event (h, t).

Let R and S be two random variables and $c \in \mathbb{R}$. The following properties are satisfied:

$$\mathbb{E}[c] \coloneqq c \tag{1.3}$$

$$\begin{aligned}
\mathbb{E}[R + c] &\coloneqq \sum_{a \in \mathcal{A}} (\mathbb{P}[R = a] \cdot (a + c)) \\
&= \sum_{a \in \mathcal{A}} (\mathbb{P}[R = a] \cdot a) + \sum_{a \in \mathcal{A}} (\mathbb{P}[R = a] \cdot c) \\
&= \mathbb{E}[R] + c \cdot \sum_{a \in \mathcal{A}} \mathbb{P}[R = a] \\
&= \mathbb{E}[R] + c \cdot 1 = \mathbb{E}[R] + c
\end{aligned} \tag{1.4}$$

$$\begin{aligned}
\mathbb{E}[R + S] &\coloneqq \sum_{a \in (\mathcal{A}_R \cup \mathcal{A}_S)} \mathbb{P}[R = a] \cdot a \\
&= \sum_{a \in \mathcal{A}_R} \mathbb{P}[R = a] \cdot a + \sum_{a \in \mathcal{A}_S} \mathbb{P}[R = a] \cdot a \\
&= \mathbb{E}[R] + \mathbb{E}[S]
\end{aligned} \tag{1.5}$$

$$\begin{aligned}
\mathbb{E}[c \cdot X] &\coloneqq \sum_{a \in \mathcal{A}} \mathbb{P}[R = a] \cdot (c \cdot a) \\
&= c \cdot \sum_{a \in \mathcal{A}} \mathbb{P}[R = a] \cdot a = c \cdot \mathbb{E}[X]
\end{aligned} \tag{1.6}$$

1.1.4 Continous probability model

Let f be a continuous function defined in $(-\infty, \infty) \subseteq \mathbb{R}$. If f satisfies the following properties, f is called a Probability Density Function (PDF):

$$f(x) \geq 0 \qquad \forall x \in (-\infty, \infty) \tag{1.7}$$

$$1 = \int_{-\infty}^{\infty} f(x)\, dx \tag{1.8}$$

Probability is defined as follows:

$$\mathbb{P}[a \leq R \leq b] \coloneqq \int_{a}^{b} f(x)\, dx \tag{1.9}$$

The *expected value* in the continuous model is defined as,

$$\mathbb{E}[R] \coloneqq \int_{-\infty}^{\infty} (\mathbb{P}[R = x] \cdot x)\ dx \coloneqq \int_{\mathbb{R}} (\mathbb{P}[R = x] \cdot x)\ dx \tag{1.10}$$

Let R and S be two random variables and $c \in \mathbb{R}$. The expected value satisfies the following properties:

$$\mathbb{E}[c] \coloneqq c \tag{1.11}$$

$$\begin{aligned}
\mathbb{E}[R + c] &\coloneqq \int_{\mathbb{R}} \mathbb{P}[R = x] \cdot (x + c)\, dx \\
&= \int_{\mathbb{R}} (\mathbb{P}[R = x] \cdot x + \mathbb{P}[R = x] \cdot c)\ dx \\
&= \int_{\mathbb{R}} \mathbb{P}[R = x] \cdot x\, dx + c \cdot \int_{\mathbb{R}} \mathbb{P}[R = x]\, dx \\
&= \mathbb{E}[R] + c \cdot 1 = \mathbb{E}[R] + c
\end{aligned} \tag{1.12}$$

$$\begin{aligned}
\mathbb{E}[R + S] &\coloneqq \int_{\mathbb{R}} (\mathbb{P}[R = x] \cdot x + \mathbb{P}[S = x] \cdot x)\ dx \\
&= \int_{\mathbb{R}} \mathbb{P}[R = x] \cdot x\, dx + \int_{\mathbb{R}} \mathbb{P}[S = x] \cdot x\, dx \\
&= \mathbb{E}[R] + \mathbb{E}[S]
\end{aligned} \tag{1.13}$$

$$\begin{aligned}
\mathbb{E}[c \cdot X] &\coloneqq \int_{\mathbb{R}} \mathbb{P}[R = x] \cdot (x \cdot c)\, dx \\
&= c \cdot \int_{\mathbb{R}} \mathbb{P}[R = x] \cdot x\, dx = c \cdot \mathbb{E}[X]
\end{aligned} \tag{1.14}$$

Because the expected value operator $\mathbb{E}$ satisfies the same properties in the discrete and continuous case, we combine those cases and denote $\mathbb{E}[R]$ for both cases for a random variable R.

1.1.5 Variance and standard deviation

Variance quantifies how strong values are spread out from their population mean:

$$\sigma^2 = \mathbb{V}[R] \coloneqq \mathbb{E}[(R - \mu)^2]$$

In the discrete case, this is equivalent to,

$$\mathbb{V}[R] = \mathbb{E}\left[\sum_{a \in \mathcal{A}} (\mathbb{P}[R = a] \cdot (a - \mu))^2\right]$$

and in the continuous case, we have:

$$\mathbb{V}[R] = \mathbb{E}\left[\int_{\mathbb{R}} (\mathbb{P}[R = x] \cdot (x + \mu))^2\ dx\right]$$

The *standard deviation* is defined as its root:

$$\sigma = \sqrt{\mathbb{V}[R]}$$

1.1.6 Law of Large Numbers

The Law of Large Numbers stresses the practical importance of the expected value. First, we define the notion of the average value over a sample A_i of size n:

$$\overline{R}_n \coloneqq \frac{1}{n}\sum_{i=0}^{n} A_i$$

Then, the Law of Large Numbers states that,

$$\lim_{n\to\infty} \overline{R}_n = \mathbb{E}[R]$$

In order to prove this theorem, we use Chebyshev's Inequality, the Weak Law of Large numbers and the Borel-Cantelli Lemma.

Consider the continuous case. Let f be a PDF, $a \in \mathbb{R}_{\geq 0}$, $n \in \mathbb{N}$ and let $\mathbb{E}[R^n]$ be defined as follows:

$$\begin{aligned}
\mathbb{E}[R^n] &= \int_{\mathbb{R}} x^n \cdot f(x)\, dx \\
&\geq \int_{x\geq a} x^n \cdot f(x)\, dx \\
&\geq a^n \int_{x\geq a} f(x)\, dx \\
&= a^n \cdot \mathbb{P}[X \geq a]
\end{aligned}$$

The discrete case follows immediately. This concludes the correctness of the following theorem:

Theorem 1 (Chebyshev's Inequality Theorem) Let R be a random variable, $a \in \mathbb{R}_{\geq 0}$ and $n \in \mathbb{N}$. Then it holds that,

$$\mathbb{P}[R \geq a] \leq \frac{1}{a^n}\mathbb{E}[R^n]$$

The next theorem is called Weak Law of Large Numbers.

Theorem 2 (Weak Law of Large Numbers, Bernoulli's Theorem) Let R_i be a sequence of independent and identically distributed random variables with common mean μ and variance σ^2. Let

$$S_n \coloneqq \sum_{i=1}^{n} R_i \qquad\qquad S_n^* \coloneqq \frac{S_n}{n} - \mu$$

Then for any $\varepsilon > 0$,

$$\mathbb{P}[S_n^* \geq \varepsilon] \leq \frac{1}{\varepsilon^2}\mathbb{V}[S_n^*]$$

TODO: iid should/will be introduced in section 1.2, but is used here already.

Proof 1

$$\begin{aligned}\mathbb{E}[S_n^*] &= \mathbb{E}\left[\frac{1}{n}\left(R_1 + R_2 + \ldots + R_n\right) - \mu\right] \\ &= \frac{1}{n}\left(\mathbb{E}[R_1] + \mathbb{E}[R_2] + \ldots + \mathbb{E}[R_n]\right) - \mathbb{E}[\mu] \\ &= \frac{n \cdot \mu}{n} - \mu = 0 \qquad (1.15)\\ \mathbb{V}[S_n^*] &= TODO \qquad (1.16)\end{aligned}$$

1.1.7 Union and intersection of events

TODO

1.1.8 Marginalization

TODO

1.1.9 Independence

TODO

1.1.10 Conditional independence

TODO definition
TODO mutual independence $\neq$ pairwise independence

1.1.11 Bayes' Theorem

Theorem 3 (Bayes' Theorem) Let A and B be two events and $\mathbb{P}[B] \neq 0$. Then:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A] \cdot \mathbb{P}[A]}{\mathbb{P}[B]}$$

Proof 2 In section 1.1.9, we showed the following relation between marginal and conditional probability:

$$\mathbb{P}[A, B] = \mathbb{P}[B|A] \cdot \mathbb{P}[A] = \mathbb{P}[A|B] \cdot \mathbb{P}[B]$$

Bayes' theorem follows immediately:

$$\frac{\mathbb{P}[B|A] \cdot \mathbb{P}[A]}{\mathbb{P}[B]} = \mathbb{P}[A|B]$$

Bayes' Theorem is fundamental to theory we will cover in the following. $\mathbb{P}[A]$ is called *prior probability* and $\mathbb{P}[A|B]$ is called *posterior probability* in the Bayesian interpretation. The names derive from the fact, that $\mathbb{P}[A]$ is known beforehand in most applications and $\mathbb{P}[A|B]$ is the degree of belief in A after B happened.

1.2 Probability distributions

TODO

1.3 Graphical models

Show symmetry, decomposition, weak union and contraction, via Dawid et al.

1.4 Example: Polynomial curve fitting problem

1.4.1 The problem

In the following, we introduce the curve fitting problem similar to [1, p. 4 ff.]. The problem is defined as follows:

Problem 1 (Polynomial curve fitting problem) Consider a polynomial of arbitrary degree.

Given $x = (x_1, \ldots, x_n)^N$ as a vector of N x-values and $t = (t_1, \ldots, t_n)^N$ as the corresponding y-values drawn from the polynomial. Furthermore let $E(w)$ be an error function for given polynomial coefficients w.

Find a polynomial with coefficients w which approximates values t minimizing $E(w)$.

The degree of the polynomial is unknown on purpose. *Model selection* is a branch of Machine Learning dedicated to finding appropriate models for given problems. So for polynomial degree choice for our curve fitting problem, we refer to research literature in Model Selection. Popular error functions include

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}(y(x_n, w) - t_n)^2 \qquad \text{(Mean squared error, MSE)}$$

$$E(w) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y(x_n, w) - t_n)^2} \qquad \text{(Root mean square, RMS)}$$

$$E(w) = \frac{1}{N}\sum_{n=1}^{N}(y(x_n, w) - t_n) \qquad \text{(Mean signed deviation, MSD)}$$

1.4.2 Overfitting

Machine Learning distinguishes between a *training* and *validation* dataset as input. It uses the training set to learn which output is desired for some given input. Therefore all elements of the training set are labelled such that the error in the output can be quantified. *Overfitting* describes the situation, when the learning algorithm approximates the output with little error, but input from the validation set (which contains different inputs) is computed with high error.

TODO: visualization

1.4.3 Regularization as countermeasure

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}(y(x_n, w) - t_n)^2 + \frac{\lambda}{2}|w|^2$$

We now model the problem from a probabilistic view:

1.4.4 Maximum Likelihood Estimator

The Maximum Likelihood Estimator (MLE) is a technique to estimate the parameters of a probability distribution. It maximizes the likelihood that the given data actually occurs.

Theorem 4 Consider input data x, mean μ and variance σ^2:

$$\ln \mathbb{P}[x|\mu, \sigma^2] = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

Then $\mu_{\text{ML}} = \frac{1}{N}\cdot\sum_{n=1}^{N} x_n$ for maximized μ and
$\sigma_{\text{ML}} = \frac{1}{N}\cdot\sum_{n=1}^{N}(x_n - \mu_{\text{ML}})^2$ for maximized σ^2

So we want to determine the 2 parameters of a Gaussian distribution, namely μ and σ^2, in the maximum likelihood case. We begin with μ:

Proof 3

1. Derive $\ln \mathbb{P}[x|\mu, \sigma^2]$ for μ

$$\begin{aligned}
\frac{\partial}{\partial\mu}\ln \mathbb{P}[x|\mu, \sigma^2] &= \frac{\partial}{\partial\mu}\left(-\frac{1}{2\sigma^2}\cdot\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)\right) \\
&= \frac{\partial}{\partial\mu}\left(-\frac{1}{2\sigma^2}\cdot\sum_{n=1}^{N}(x_n^2 - 2x_n\mu + \mu^2) - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)\right) \\
&= -\frac{1}{2\sigma^2}\cdot\sum_{n=1}^{N}(-2x_n + 2\mu) \\
&= -\frac{1}{\sigma^2}\cdot\sum_{n=1}^{N}(\mu - x_n)
\end{aligned}$$

2. Set result zero

$$0 = -\frac{1}{\sigma^2}\cdot\sum_{n=1}^{N}(\mu - x_n) = \sum_{n=1}^{N}(\mu - x_n) = N\cdot\mu - \sum_{n=1}^{N} x_n$$

$$\implies \mu_{\text{ML}} = \frac{1}{N}\cdot\sum_{n=1}^{N} x_n \qquad \text{commonly called “sample mean”}$$

We continue with σ^2 and use the same approach:

Proof 4 1. Derive $\ln \mathbb{P}[x|\mu,\sigma^2]$ for σ^2

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\ln\mathbb{P}[x|\mu,\sigma^2] &= \frac{\partial}{\partial\sigma^2}\left(-\frac{1}{2\sigma^2}\cdot\sum_{n=1}^{N}(x_n-\mu)^2-\frac{N}{2}\ln\sigma^2-\frac{N}{2}\ln(2\pi)\right)\\ &= \frac{1}{2\sigma^4}\cdot\sum_{n=1}^{N}(x_n-\mu)^2-\frac{N}{2}\cdot\frac{1}{\sigma^2}\\ &= \frac{1}{2\sigma^2}\left(\frac{1}{\sigma^2}\cdot\sum_{n=1}^{N}(x_n-\mu)^2-N\right)\end{aligned}$$

2. Set result zero

$$\begin{aligned}0 &= \frac{1}{2\sigma^2}\left(\frac{1}{\sigma^2}\cdot\sum_{n=1}^{N}(x_n-\mu)^2-N\right)\\ N\cdot\sigma^2 &= \sum_{n=1}^{N}(x_n-\mu)^2\\ \sigma^2_{\text{ML}} &= \frac{1}{N}\cdot\sum_{n=1}^{N}(x_n-\mu)^2 \qquad \text{commonly called “sample variance”}\end{aligned}$$

And now we derive the precision parameter β in the maximum likelihood case:

Theorem 5 Given

$$\ln\mathbb{P}[t|x,w,\beta] = -\frac{\beta}{2}\cdot\sum_{n-1}^{N}\left(y(x_n,w)-t_n\right)^2+\frac{N}{2}\ln\beta-\frac{N}{2}\ln(2\pi)$$

then find

$$\frac{1}{\beta_{\text{ML}}} = \frac{1}{N}\cdot\sum_{n=1}^{N}(y(x_n,w_{\text{ML}})-t_n)^2$$

by maximizing β

Proof 5 1. Derive $\ln\mathbb{P}[t|x,w,\beta]$ with β

$$\begin{aligned}\frac{\partial}{\partial\beta}\ln\mathbb{P}[t|x,w,\beta] &= \frac{\partial}{\partial\beta}\left(-\frac{\beta}{2}\sum_{n=1}^{N}(y(x_n,w)-t_n)^2+\frac{N}{2}\ln\beta-\frac{N}{2}\ln(2\pi)\right)\\ &= -\frac{1}{2}\cdot\sum_{n=1}^{N}\left(y(x_n,w)-t_n\right)^2+\frac{N}{2}\cdot\frac{1}{\beta}\end{aligned}$$

2. Set result zero

$$\begin{aligned}0 &= -\frac{1}{2}\cdot\sum_{n=1}^{N}\left(y(x_n,w)-t_n\right)^2+\frac{N}{2\beta}\\ \frac{N}{\beta} &= \sum_{n=1}^{N}\left(y(x_n,w)-t_n\right)^2\\ \frac{1}{\beta_{\text{ML}}} &= \frac{1}{N}\cdot\sum_{n=1}^{N}(y(x_n,w)-t_n)^2\end{aligned}$$

Chapter 2

Neural networks

Bibliography

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. isbn: 0387310738.