Advanced probabilistic machine learning

Linus Falk

September 3, 2023

Contents

1 Introduction

This course bla bla bla

2 Probability review

To begin with, lets start with a review of what we already know about probability and a look at the *Sample space*:

- Dice: $\Omega = \{1,2,3,4,5,6\}$
- Coin flip: $\Omega = \{\text{heads, tails}\}$

There are also *Events* which is a subset of the sample space:

- Dice: A = $\{1,2\}$
- Coin flip = $\{\text{tails}\}$

Three important axioms to remember are.

1. P(A) ≥ 0
2. P(Ω) $= 1$
3. For disjoint sets, A_1,A_2, ...
   $P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$

To sum it up here are some consequences of the these axioms:

1. $A \subseteq B \Rightarrow P(A) \leq P(B)$
2. $P(A^c) = 1 - P(B)$
3. $P(\varnothing) = 0$
4. $0 \leq P(A) \leq 1$
5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

2.1 Conditional probability

The definition of conditional probability for two events are the following:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{1}$$

Provided that the probability of event A is greater than 0, $P(A) > 0$. From this we can derive the **product rule:** $P(A \cap B) = P(B|A)P(A)$

2.2 Bayes theorem

Given two events called A and B:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{2}$$

Example: medical test paradox

Let's characterize a medical test:

- Sensitivity(True positive rate) $= P(+|\text{Disease})$
- Specificity (True negative rate) $= P(-|\text{No Disease})$

Then characterize a medical condition in a population:

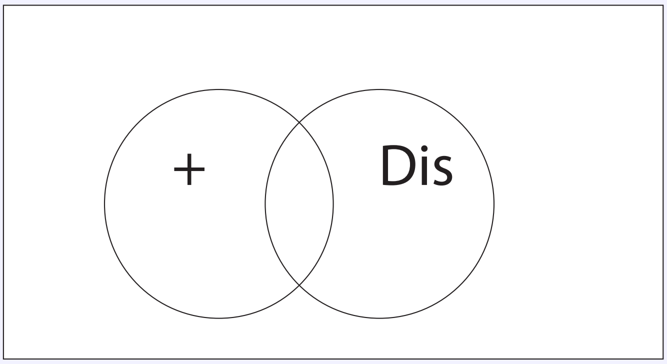
- Prevalence = "*proportion of a particular population found to be affected by a medical condition.*"

Question: Assume that for a certain disease the Sensitivity is 0.99, Specificity is 0.99 and Prevalence = 0.01 If you pick someone randomly from the population and test the person for this disease, obtaining a positive result. What is the probability of the person actually having the disease?

Using Bayes theorem:

$$P(\text{Disease}|+) = \frac{P(+|\text{Disease})P(\text{Disease})}{P(+)} = \frac{\text{tpr} \cdot p}{P(+)} \tag{3}$$

and by the law of total probability:
$P(+) = P(+ \cap \text{Disease}) + P(+ \cap \text{No disease})$
$P(+) = P(+|\text{disease})P(\text{disease}) + (1 - P(-|\text{No disease}))(1 - P(disease))$
$P(+) = \text{tpr} \cdot p + (1 - \text{tnr})(1 - p)$



Hence we have:

$$P(\text{disease}|+) = \frac{0.99 \cdot 0.001}{0.99 \cdot 0.01 + 0.001 \cdot 0.99} \approx \frac{1}{11} \tag{4}$$

2.3 Random variables

The definition of a random variable $\mathcal{X}$ is a quantity that depends on a random event. Where a discrete random variable is a countable number of values. Described by the probability mass function $p(x) = P(X = x)$. An **example:** 6-sided dice with $p(x) = 1/6, x = 1, \ldots, 6$. A continuous random variable is an uncountable number of values. Ex: any value in $\mathbb{R}$. **Example:** $\mathcal{X}$ is the height of an adult randomly sampled from the population. It is described by the probability density function: $p(x)$.
A probability density function $p(x)$ describes the probability for a *continuous* random variable x falling into a given interval:

$$P(a < X < b) = \int_a^b p(x)\, \mathrm{d}x \tag{5}$$

2.4 Joint probability

Given two random variables: X and Y:

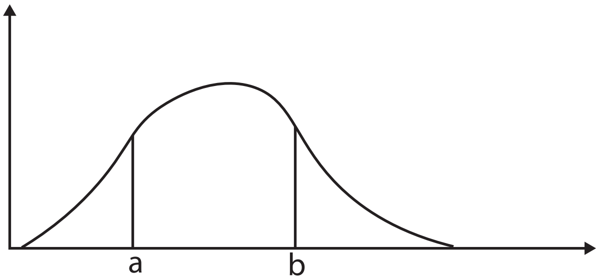


Figure 1: Example of caption

- Discrete: probability mass function
  $p(x, y) = P(X = x, Y = y)$
- Continuous: (probability density function)
  $P(a < X \leq b, c < Y \leq d) = \int_a^b \int_c^d p(x, y)\,\mathrm{d}y\mathrm{d}x$

2.5 Marginalization and conditioning

Marginalization (also called the **sum rule**) is defined as

$$\begin{aligned} p(x) &= \int_y p(z, y)\,\mathrm{d}y \quad \text{if y is continuous} \\ p(x) &= \sum_y p(x, y) \quad \text{if y is discrete} \end{aligned} \tag{6}$$

Conditional probability also called the **product rule** is defined as:

$$\begin{aligned} p(x, y) &= \frac{p(x, y)}{p(y)} \quad \text{where} p(y) \neq 0 \\ &\Rightarrow p(x, y) = p(x|y)p(y) \end{aligned} \tag{7}$$

2.6 Bayes theorem: random variables

Both for probability mass functions and probability density functions:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \tag{8}$$

2.7 Probabilistic modelling

In probabilistic modelling we consider two types of variables:

$$\begin{aligned} \mathcal{D} = \{x_1, x_2, \ldots, X_N\}: &\quad \text{observed variables} \\ \Theta = \{x_1, x_2, \ldots, X_N\}: &\quad \text{latent variables} \end{aligned} \tag{9}$$

In **probabilistic modelling** we treat *both* the observed variables and the latent variables as **random variables**

We model this relationship between $\mathcal{D}$ and Θ with its **Joint distribution**

$$p(\mathcal{D}, \Theta) = p(x_1, \ldots, X_N, z_1, \ldots, z_M) \tag{10}$$

2.8 Bayes theorem in Probabilistic modelling

The joint distribution factorizes into **likelihood** and a **prior**

$$p(\mathcal{D}, \Theta) = p(\mathcal{D}|\Theta)p(\Theta) \tag{11}$$

We can now find $p(\Theta|\mathcal{D})$ using **Bayes' theorem**

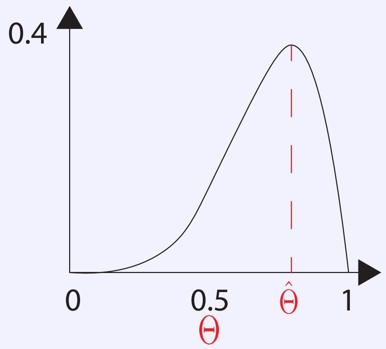
$$p(\Theta|\mathcal{D}) = \frac{p(\mathcal{D}|p(\Theta))}{p(\mathcal{D})} \tag{12}$$

- $\mathcal{D}$: observed data
- Θ : latent variables explaining the data
- $p(\Theta)$: **prior** belief of latent variables before seeing data
- $p(\mathcal{D}|\Theta)$: **likelihood** of the data in view of the latent variables
- $p(\Theta|\mathcal{D})$: **posterior** belief of latent variables in view of data
- $p(\mathcal{D})$: The **marginal likelihood** (presented in lecture 3)

Example: Coin flip (>Frequentist viewpoint)

- $x \in \{0, 1\}$ represent the outcome of flipping a damaged coin
- $p(x = 1|\Theta) = \Theta$, is a deterministic parameter
- Nor prior belief encoded

Question: After we observe N coin flips $\mathcal{D} = \{x_1, \ldots, x_N\}$, which $\hat{\Theta}$ makes the observed data most likely?
Solution: Maximize the likelihood function $p(\mathcal{D}|\Theta)$



Example: Coin flip (Bayesian viewpoint)

- $x \in \{0, 1\}$ represent the outcome of flipping a damaged coin
- $p(x = 1|\mu) = \mu$, which is also a random variable.
- Probability distribution $p(\mu)$: our *prior belief*

Question: After we observe N coin flips $\{x_1, \ldots, x_N\}$, what is our belief $p(\mu|x_1, \ldots, x_N)$?
Solution: Bayes theorem states that:

$$p(\mu|x_1, \ldots, x_N) \propto p(x_1, \ldots, x_N|\mu)p(\mu) \tag{13}$$

We will continue with this example next lecture...

2.9 Condluding remarks

Probabilistic/Bay inference is a flexible way of dealing the machine learning problems. Some properties to remember:

- Treat not only the data, but also the model and its parameters (if they are parametric) as random variables
- After learning you not only get a single model, you get a distribution of likely models.
- You can also encode prior knowledge you might have about the model and its parameters.

2.10 Summary

Probability distribution is a function that describes the likelihood of obtaining the possible values that a random variable can assume. **Conditional and marginalization** are two basic rules for manipulating probability distributions. **Frequentist vs Bayesian**: the first assume true values underlying some experiment and the second require some initial belief to be set on possible values.

- **Bayes theorem**, $p(x|y) = p(y|x)p(x)/p(y)$
- **Prior**, belief of parameters before we have seen any data
- **Likelihood**, belief of data in view of the parameters
- **Posterior**, belief of parameters after inferring data

3 Bayes' theorem

We can find $p(\Theta|\mathcal{D})$ by using **Bayes' theorem**

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\Theta)p(\Theta)}{p(\mathcal{D})} \tag{14}$$

We can view $p(\mathcal{D})$ as a normalized constant and if we view the quantities as functions of Θ we can write:

$$\underset{\text{posterior}}{p(\Theta|\mathcal{D})} \propto \underset{\text{likelihood}}{p(\mathcal{D}|\Theta)}\,\underset{\text{prior}}{p(\Theta)} \tag{15}$$

$\propto$ means that it is proportional with respect to Θ

Binomial-beta conjugate pair

The binomial-beta conjugate pair refers to the Bayesian statistical framework where the binomial distribution, describing the likelihood of observed successes and failures, is combined with a beta distribution prior, resulting in a beta distribution posterior.

Example: Flipping a damaged coin

- x = 1 represents "head" and x = 0 represents "tail"
- The probability of x = 1 is: $p(x=1|\mu) = \mu, \quad 0 \leq \mu \leq 1$

we assume a damaged coin here, so it is not necessary 0.5

Question: Given a dataset $\mathcal{D} = \{x_1, \dots, x_n\}$, what is $p(\mu|\mathcal{D})$?

Solution: Bayes' theorem states that

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})} \tag{16}$$

Find the likelihood and prior and then multiply! Lets start with the likelihood
We know that :

$$\begin{aligned} p(x=1|\mu) &= \mu \\ p(x=1|\mu) &= 1-\mu \end{aligned} \tag{17}$$

which is known as the Bernoulli distribution

$$p(x|\mu) = \mathrm{Bern}(x;\mu) = \mu^x(1-\mu)^{1-x} \tag{18}$$

The observations $x_1, \dots, x_N$ are conditionally independent given μ and the likelihood is then:

$$\begin{aligned} p(x_1, \dots, x_N|\mu) &= \prod_{n=1}^{N} p(x_n|n) \\ &= \prod_{n=1}^{N} p(x_n|n)\mu^{x_n}(1-\mu)^{1-x_n} \\ &= \mu^m(1-\mu)^{N-m} \end{aligned} \tag{19}$$

where $m = \sum_{n=1}^{N} x_n$ i.e the number of heads. **Note**: the likelihood only depend on the data $\mathcal{D}$ via m. Hence we can modify our problem slightly $p(\mu|m) \propto \underbrace{p(m|\mu)}_{\text{likelihood}}\underbrace{p(\mu)}_{\text{prior}}$
Hence the likelihood is given by the binomial distribution

$$p(m|\mu) = \mathrm{Bin}(m;N,\mu) = \binom{N}{M}\mu^m(1-\mu)^{N-m} \tag{20}$$

where $\binom{N}{m} = \frac{N!}{(N-m)!m!}$ is the number of sequences giving m heads.

Remember Bayes theorem $p(\mu|m) \propto p(m|\mu)p(\mu)$. There exists multiple possible prior distributions $p(\mu)$ and we opt for a prior which has attractive analytical properties. So we choose a prior such that the posterior will be of the same functional form as the prior. We call this a conjugate prior. The conjugate prior of the Binomial distribution is the Beta distribution:

$$\begin{aligned} \mathrm{Beta}(\mu;a,b) &\propto \mu^{a-1}(1-\mu)^{b-1} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\ \text{where } \Gamma(a) &= \int_0^{\infty} e^{-x}x^{a-1}\,\mathrm{d}x \end{aligned} \tag{21}$$

The posterior can now be computed

$$\begin{aligned} p(\mu|m) &\propto p(m|\mu)p(\mu) \\ &= \mathrm{Bin}(m;N,\mu)\mathrm{Beta}(\mu;a,b) \\ &\propto \mu^m(1-\mu)^{N-m}\mu^{a-1}(1-\mu)^{b-1} \\ &= \mu^{m+a-1}(1-\mu)^{N-m+b-1} \end{aligned} \tag{22}$$

Example: cont.

Hence is the posterior also a Beta distribution:

$$p(\mu|m) = \text{Beta}(\mu; a^\star, b^\star)$$
$$\text{where} a^\star = m + 1, b^\star = N - m + b \tag{23}$$

Bayesian inference:

- Start with equal probability for all $\mu \Rightarrow \text{Beta}(\mu; 1, 1) = 1$
- Assume that we get one data point $x_1 = 1$
- The posterior $\propto$ likelihood $\times$ prior

Continue and assume we get new data points, $N = 5$ of which $m = 4$ are heads, $\mathcal{D} = \{1, 0, 1, 1, 1\}$.
See slides, update later

Real world applications

To give an idea of then this can be used, here are some real world applications:

- **Engineering**: Estimating the proportion of aircraft turbines blades that posses a structural defect after fabrication.
- **Social science**: Estimating the proportion of individuals who respond yes on a census question.
- **Data science**: Estimating the proportion of individuals who click on an ad when visiting a website

4 Gauss-Gauss conjugate pair

The Gauss-Gauss conjugate pair refers to the Bayesian statistical framework where a Gaussian (normal) distribution likelihood, describing observed data, is combined with a Gaussian prior, resulting in a Gaussian posterior.
So for a scalar random variable X:

- Expected value:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x \tag{24}$$

- Variance:

$$\text{Var}[X] = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - \mu^2, \quad \text{where } \mu = \mathbb{E}[X] \tag{25}$$

- Standard deviation: $\alpha = sqr^{\text{Var}[X]}$

For a scalar variable x. the Gaussian distribution can be written on the form:

$$\mathcal{N}(x; \hat{\mu}, \Sigma) = \underbrace{\frac{1}{(2\pi)^{D/2}\sqrt{\det \Sigma}}}_{Z} \exp\left(-\frac{1}{2}(x - \hat{\mu}^T \Sigma^{-1}(x - \hat{\mu})\right) \tag{26}$$

- $\mathbb{E}[x] = \hat{\mu}$
- $\text{Cov}[x] = \Sigma$
- item Z is the normalization constant

We let:

$$p(x) = \mathcal{N}(x; \mu, \sigma^2)$$
$$p(y|x) = \mathcal{N}(y; x, \eta^2) \tag{27}$$

We then have:

$$
\begin{aligned}
p(x|y) &= \frac{p(y|x)p(x)}{p(y)} \\
&= \mathcal{N}(x; m, s^2)
\end{aligned}
\tag{28}
$$

with the following:

$$
\begin{aligned}
s^2 &= \frac{1}{\sigma^{-2} + \eta^{-2}} \\
m &= \frac{\sigma^{-2} + \eta^{-2} y}{\sigma^{-2} + \eta^{-2}}
\end{aligned}
\tag{29}
$$

Expected value vector

For a vector random variable $\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix}$ we have:

- Expected value:

$$
\begin{bmatrix} \mathcal{E}[X_1] \\ \vdots \\ \mathcal{E}[X_N] \end{bmatrix}
\tag{30}
$$

- The covariance matrix:

$$
\text{Cov}[\bar{X}] = \mathcal{E}[(\bar{X} - \bar{\mu})(\bar{X} - \mu)^T] = \mathbb{E}[\mathbf{X}\mathbf{X}^T] - \bar{\mu}\bar{\mu}^T
\tag{31}
$$

Multivariate Gaussian

For a D-dimensional vector x, the multivariate Gaussian distribution can be written on the form:

$$
\mathcal{N}(x; \bar{\mu}, \Sigma) = \mathcal{N}(x; \hat{\mu}, \Sigma) = \underbrace{\frac{1}{(2\pi)^{D/2}\sqrt{\det\Sigma}}}_{Z} \exp\left(-\frac{1}{2} \underbrace{(x - \hat{\mu})^T \Sigma^{-1} (x - \hat{\mu})}_{\text{quadtratic form}} \right)
\tag{32}
$$

- $\mathbb{E}[x] = \mu$
- $\text{Cov}[x] = \Sigma$
- Z is the normalization constant

The Gaussian in proportional to $e^{\text{quadtratic form}}$

Partitioned Gaussian

Partition the Gaussian random vector $\bar{x} \sim \mathcal{N}(\bar{\mu}, \Sigma)$, where $\bar{x} \in \mathbb{R}^n$ into two sets of random variables $\bar{x_a} \in \mathbb{R}^{n_a}$ and $\bar{x_b} \in \mathbb{R}^{n_b}$,

$$
x = \begin{pmatrix} \bar{x_a} \\ \bar{x_b} \end{pmatrix}, \quad \bar{\mu} = \begin{pmatrix} \bar{\mu_a} \\ \bar{\mu_b} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}
\tag{33}
$$

Affine transformation of multivariate Gauss

Corollary 1 (Affine transformation - conditional)

If we assume that $\bar{x_a}$ as well as $\bar{x_b}$ conditioned on $\bar{x_a}$ are Gaussian distributed according to:

$$\begin{aligned} p(\bar{x_a}) &= \mathcal{N}(\bar{x_a}; \bar{\mu_a}, \Sigma_a), \\ p(\bar{x_a}|\bar{x_b}) &= \mathcal{N}(\bar{x_b}; \mathbf{A}\bar{x_a} + \mathbf{b}, \Sigma_{b|a}) \end{aligned} \tag{34}$$

Then the conditional distribution of $\bar{x_a}$ given $\bar{x_b}$ is:

$$p(\bar{x_a}|\bar{x_b}) = \mathcal{N}(\bar{x_a}; \bar{\mu_{a|b}}, \Sigma_{a|b}) \tag{35}$$

Together with:

$$\begin{aligned} \bar{\mu_{a|b}} &= \Sigma_{a|b}(\Sigma_a^{-1}\mu_a + \mathbf{A}^T\Sigma_{a|b}^{-1}(\bar{x_b} - \mathbf{b})) \\ \Sigma_{a|b} &= (\Sigma_a^{-1} + \mathbf{A}^T\Sigma_{b|a}^{-1}\mathbf{A})^{-1} \end{aligned} \tag{36}$$

Corollary 2 (Affine transformation - Marginalization)

If we assume that $\bar{x_a}$ as well as $\bar{x_b}$ conditioned on $\bar{x_a}$ are Gaussian distributed according to:

$$\begin{aligned} p(\bar{x_a}) &= \mathcal{N}(\bar{x_a}; \bar{\mu_a}, \Sigma_a), \\ p(\bar{x_a}|\bar{x_b}) &= \mathcal{N}(\bar{x_b}; \mathbf{A}\bar{x_a} + \mathbf{b}, \Sigma_{b|a}) \end{aligned} \tag{37}$$

Then the marginal distribution of $\bar{x_b}$ is then given by $\bar{x_b}$ is:

$$p(\bar{x_b}|\bar{x_b}) = \mathcal{N}(\bar{x_b}; \bar{\mu_b}, \Sigma_b) \tag{38}$$

Together with:

$$\begin{aligned} \bar{\mu_b} &= \mathbf{A}\bar{\mu_a} + \mathbf{b} \\ \Sigma_b &= (\Sigma_{b|a} + \mathbf{A}\Sigma_a\mathbf{A})^T \end{aligned} \tag{39}$$

Gaussian inference (scalar example) from Corollary 1

We let:

$$\begin{aligned} p(x) &= \mathcal{N}(x; \mu, \sigma^2) \\ p(y|x) &= \mathcal{N}(y; x, \eta^2) \end{aligned} \tag{40}$$

With:

$$\begin{aligned} \boldsymbol{x_a} = x, \quad \boldsymbol{\mu_a} = \mu, \quad \Sigma_a = \sigma^2 \\ \boldsymbol{x_b} = y, \quad \mathbf{A} = 1, \quad b\mathbf{b} = 0, \quad \Sigma_{b|a} = \eta^2 \end{aligned} \tag{41}$$

Now with Corollary 1 this gives us:

$$p(x|y) = \mathcal{N}(x; m, s^2) \tag{42}$$

where

$$\begin{aligned} s^2 &= \frac{1}{\sigma^{-2} + \eta^{-2}} \\ m &= \frac{\sigma^{-2}\mu 0\eta^{-2}y}{\sigma^{-2} + \eta^{-2}} \end{aligned} \tag{43}$$

$$
\begin{aligned}
p(x|y) &\propto p(y|x)p(x) \\
&\propto e^{-\frac{-(y-x)^2}{2\eta^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\
&\text{complete the squares} \\
&\frac{y^2-2x+x^2}{\eta^2}+\frac{x^2-2x\mu+\mu^2}{\sigma^2} \\
&= (\frac{1}{\eta^2}+\frac{1}{\sigma^2})x^2-2(\frac{y}{\eta^2}+\frac{\mu}{\sigma^2})x+\frac{y^2}{\eta^2}+\frac{\mu^2}{\sigma^2} \\
&= a(x-\frac{b}{a})^2-\frac{b^2}{a}+c \\
&e^{\frac{1}{2}a(x-\frac{b}{a})^2} e^{-\frac{b^2}{a}+c}
\end{aligned}
\tag{44}
$$

Summarize lecture 2 with a few concepts

Conjugate prior is a propr ensuring that the posterior and the prior belong to the same probability distribution family. Bernoulli distribution is a distribution for binary random variables. Binomial distribution is a distribution for the sum of multiple binary random variables.Beta distribution, the conjugate prior for the binomial distribution. Gaussian distribution, the well known probability density function with the shape of the bell curve. Multivariate Gaussian distribution is a generalization of the Gaussian distribution to higher dimensions.

Theorem 1 - Marginalization

Partition the Gaussian random vector $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ according to:

$$
\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_a \\ \boldsymbol{x_b} \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_a} \\ \boldsymbol{\mu_b} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \tag{45}
$$

The marginal distribution is then given by: $p(\boldsymbol{x}_a) = \mathcal{N}(\boldsymbol{x}_a; \boldsymbol{\mu}_a, \Sigma_{aa})$

Theorem 2: Conditioning

Partition the Gaussian random vector $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ according to:

$$
\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_a \\ \boldsymbol{x_b} \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_a} \\ \boldsymbol{\mu_b} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \tag{46}
$$

The conditional distribution $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$ is then given by:

$$
\begin{aligned}
p(\boldsymbol{x}_a|\boldsymbol{x}_b) &= \mathcal{N}(\boldsymbol{x}_a; \boldsymbol{\mu}_{a|b}, \Sigma_{a|b}), \\
\boldsymbol{\mu} &= \boldsymbol{\mu}_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\boldsymbol{x_b} - \boldsymbol{\mu}_b) \\
\Sigma_{a|b} &= \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}
\end{aligned}
\tag{47}
$$

Theorem 3 - Affine transformation

Lets assume that $\mathbf{x}_a$ as well as $\mathbf{x}_b$ conditioned on $\mathbf{x}_a$ are Gaussian distributed according to following:

$$\begin{aligned} p(\boldsymbol{x}_a) &= \mathcal{N}(\boldsymbol{x}_a; \boldsymbol{\mu}, \Sigma_a), \\ p(\boldsymbol{x}_b|\boldsymbol{x}_a) &= \mathcal{N}(\boldsymbol{x}_b; \mathbf{A}\boldsymbol{x}_a + \mathbf{b}, \Sigma_{b|a}) \end{aligned} \tag{48}$$

The joint distribution of $\mathbf{x}_a$ and $\mathbf{x}_b$ is then:

$$p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_a \\ \mathbf{A}\boldsymbol{\mu}_a + \mathbf{b} \end{bmatrix}, \mathbf{R}\right) \tag{49}$$

with R equal to:

$$\mathbf{R} = \begin{bmatrix} \Sigma_a & \Sigma_a \mathbf{A}^T \\ \mathbf{A}\Sigma_a & \Sigma_{b|a} + \mathbf{A}\Sigma_a\mathbf{A}^T \end{bmatrix} \tag{50}$$

5 Supervised machine learning

Supervised machine learning involves methods for learning a model for the relationship between the input x and the output y from some observed data set $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$. This lecture will cover how we can use what we learned in the course: Statistical machine learning into use with the Bayesian concept. We will do this by introducing: **Bayesian linear regression**

The model is after training used to **predict** the output from an unseen input. The question for us is, how to reformulate the supervised machine learning problem into the probabilistic setting?

5.1 Classic linear regression

Lets recall the linear regression form the SML course:

$$y_n = \boldsymbol{w}^T \boldsymbol{x_n} \epsilon_n, \quad \epsilon \sim \mathcal{N}(0, \Sigma^2), \tag{51}$$

- y_n - observed random variable
- **w** - unknown deterministic variable
- x_n known deterministic variable
- ϵ_n - unknown random variable
- σ - known deterministic variable

5.2 Linear regression: Maximum likelihood

There are two equivalent ways of expressing the linear regression model:

- $y_n = \boldsymbol{w}^T x_n + \epsilon_n, \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$
- $p(y_n|\boldsymbol{w}) = \mathcal{N}(y_n; \boldsymbol{w}^T \boldsymbol{x}_n, \sigma^2$

The likelihood $p(y_n|\boldsymbol{w})$ is given by:

$$p(\boldsymbol{y}|\boldsymbol{w}) = \prod_{n=1}^{N} p(y_n|w) = \prod_{n=1}^{N} \mathcal{N}(y_n; w^T x_n, \sigma^2) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{X}\boldsymbol{w}, \sigma^2 \boldsymbol{I}_N) \tag{52}$$

The solution is found by maximizing the likelihood: $\hat{w} = \arg\max_{w} p(\boldsymbol{y}|w)$

5.3 Bayesian linear regression

We now introduce a prior over the parameter w. Bayesian linear regression model:

$$\boxed{\begin{aligned} y_n = \boldsymbol{w}^T \boldsymbol{x}_n + \epsilon_n, \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \quad n = 1, \ldots, N, \\ \boldsymbol{w} \sim p(\boldsymbol{w}) \end{aligned} \tag{53}}$$

The present assumptions for the model:

- y_n - observed random variable
- **w** - unknown random variable
- x_n known deterministic variable
- ϵ_n - unknown random variable
- σ - known deterministic variable

The task is to compute the posterior distribution: $p(w|y)$.

Recall from lecture two, the multivariate Gaussian. For a D-dimensional vector x, the multivariate Gaussian distribution can be written on the form:

$$\mathcal{N}(x;\boldsymbol{\mu},\Sigma) = \underbrace{\frac{1}{(2\pi)^{D/2}\sqrt{\text{get}\Sigma}}}_{Z}\exp\left(-\frac{1}{2}(x-\boldsymbol{\mu})^T\Sigma^{-1}(x-\boldsymbol{\mu})\right) \tag{54}$$

- $\boldsymbol{\mu}$ is the mean vector
- Σ is the co variance matrix
- Z is the normalization constant

The probabilistic model is given by:

$$\begin{aligned} pp(y|w) &= \mathcal{N}(y;\boldsymbol{X}\boldsymbol{w},\beta^{-1}\boldsymbol{I}_N), \quad \text{likelihood} \\ p(w) &= \mathcal{N}(w;\boldsymbol{m}_0,S_0), \quad \text{prior distribution} \end{aligned} \tag{55}$$

The task is to compute the posterior distribution: $p(w|y)$. The solution is to identify:

$$x_a = w, \quad x_b = y \tag{56}$$

We use Corollary 1 to get the posterior distribution:

$$pp(w|y) = \mathcal{N}(w, \boldsymbol{m}_N, \boldsymbol{S}_n) \tag{57}$$

where:

$$\begin{aligned} \boldsymbol{m}_N &= \boldsymbol{S}_N(\boldsymbol{S}^{-1}\boldsymbol{m}_0 + \beta\boldsymbol{X}^T\boldsymbol{y}) \\ \boldsymbol{S}^{-1} &= \boldsymbol{S}_0^{-1} + \beta\boldsymbol{X}^T\boldsymbol{X} \end{aligned} \tag{58}$$

Example: Bayesian linear regression

Consider the problem of fitting a straight line to noisy measurements. We let the model be : $y_n \in \mathbb{R}, x_n \in \mathbb{R}$

$$y_n = w_0 + w_1 x_n + \epsilon_n), \quad n = 1, \dots, N \tag{59}$$

where:

$$\begin{gathered} x_n = \begin{bmatrix} 1 \\ x_n \end{bmatrix}, \quad w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \\ \epsilon_n \sim \mathcal{N}(w|(0,0)^T, \alpha^{-1}\boldsymbol{I}_2), \quad \text{where } \alpha = 2 \end{gathered} \tag{60}$$

- The prior : $p(w) = \mathcal{N}(w|\begin{pmatrix}0 & 0\end{pmatrix}^T, \frac{1}{2}\boldsymbol{I}_2$
- The likelihood function: $p(y_2|w) = \mathcal{N}(y_n|w_0 + w_1 x_2, \beta^{-1}$
- Posterior/prior: $p(w|y_2) = \mathcal{N}(w|\boldsymbol{m}_2,$
- $m_3 = \beta \boldsymbol{S}_3 \boldsymbol{X}^T \boldsymbol{y}$
- $\boldsymbol{S}_3 = (\alpha \boldsymbol{I}_2 + \beta \boldsymbol{X}^T \boldsymbol{X}^{-1}$

Predictive distribution, for a new data point (y_*, x_*) we have:

$$\begin{aligned} p(y_x|w) &= \mathcal{N}(y_*; \boldsymbol{x}_*^T, \beta^{-1}), \quad \text{likelihood} \\ p(w|y) &= \mathcal{N}(w; \boldsymbol{m}_N, \boldsymbol{S}_N), \quad \text{posterior} \end{aligned} \tag{61}$$

This can be applied to multiple types of regression, such as:

- Linear:

$$f(x) = \boldsymbol{x}^T \boldsymbol{w}, \quad \boldsymbol{x} = \begin{bmatrix} 1, & x \end{bmatrix}^T \tag{62}$$

- Quadratic regression:

$$f(x) = \boldsymbol{x}^T \boldsymbol{w}, \quad \boldsymbol{x} = \begin{bmatrix} 1, & x, & x^2 \end{bmatrix} \tag{63}$$

  i
- Switch regression:

$$f(x) = \boldsymbol{x}^T \boldsymbol{w}, \quad \boldsymbol{x}^T = \begin{bmatrix} h(x - S) & h(x - 6) & \dots & h(x + s) \end{bmatrix}, \quad h(x) = \tag{64}$$

5.4 Hyperparameters

We have seen a few examples of how to find the posterior for a parameter $\boldsymbol{\theta}$

$$p_\xi(\theta|\mathcal{D}) = \frac{p_\xi(\mathcal{D}|}{\boldsymbol{\theta} p_\xi(\mathcal{D})} \tag{65}$$

In probabilistic models we usually have **hyperparameters** ξ, for example:

- Coin flip example: $\xi = \{a, b\}$, the parameters of the beta prior.
- Bayesian linear regression: $\xi = \{a, b\}$, precision of prior and the likelihood.

But how to choose these hyperparameters? We have some alternatives:

1. Pick hyperparameters that reflects any prior knowledge of the problem.
2. The go-to solution for machine learning: k-fold cross validation.
3. The Bayesian alternative: Maximize the marginal likelihood.
4. And the even more Bayesian alternative: Put priors on the hyperparameters and do inference.

Since the cross validation alternative was covered in the last course we will only take a look at the two last options.

5.5 Bayesian approach 1: Marginal likelihood

The marginal likelihood/evidence $p_\xi(\mathcal{D})$ says how probable the data $\mathcal{D}$ is in view of the hyperparameters ξ. With a similar argument as for the maximum likelihood idea that was presented in the SML course, we can select ξ as :

$$\hat{\xi} = \arg\max_{\xi} p_\xi(\mathcal{D}) \tag{66}$$

Maximizing the likelihood often leads to overfit.
Maximizing the *marginal* likelihood rarely leads to overfit, (fewer parameters)

To maximize the marginal likelihood is sometimes referred to as the *emperical Bayes*. We have to use numerical optimization, such as BFGS or similar to to optimize the marginal likelihood. The **idea** is to compute the gradient $\nabla_\xi p(y)$ and numerically estimate the Hessian $\nabla_\xi^2 p(y)$. Can be done with newtons methods, one step: $\xi^{x+1} \leftarrow \xi^t - |\nabla_\xi^2 p(y)|^{-1} |\nabla_\xi p(y)|$. **Note** that it is important how you initialize the hyperparameter search.

5.6 Bayesian approach 2: Hyperpriors

The probabilistic model with the unknown w is given by:

$$\begin{aligned} p(w) &= \mathcal{N}(w; \boldsymbol{m}_0, \alpha^{-1}\boldsymbol{I}_D) \\ p(y|w) &= \mathcal{N}(y; \boldsymbol{X}\boldsymbol{w}, \beta^{-1}\boldsymbol{I}_N) \end{aligned} \tag{67}$$

which gives the posterior:

$$p(w|y) = \mathcal{N}(w; \boldsymbol{m}_N, S_N) \tag{68}$$

Note here that using a Gaussian prior gives a Gaussian posterior. Hence the Gaussian prior is a Conjugate prior for the Gaussian likelihood with an unknown w. A question we might ask here is: *What if also β precision is unknown.*

The probabilistic model with unknown w and β is given by:

$$\begin{aligned} p(w|\beta) &= \mathcal{N}(w; \boldsymbol{m}_0, \beta^{-1}\alpha^{-1}\boldsymbol{I}_N)\mathrm{Gam}(\beta; a_0, b_0), \quad \text{prior} \\ p(y|w) &= \mathcal{N}(y; \boldsymbol{X}\boldsymbol{w}, \beta^{-1}\boldsymbol{I}_N), \quad \text{likelihood} \end{aligned} \tag{69}$$

which will give us the posterior:

$$p(w, \beta|y) = \mathcal{N}(w; \boldsymbol{m}_N, \beta^{-1}\boldsymbol{S}_N)\mathrm{Gam}(\beta, a_N, b_N), \quad \text{posterior} \tag{70}$$

Using a Gauss-Gamma prior gives a Gauss-Gamma posterior. Hence the Gauss-Camma prior is a conjugate prior to the Gaussian likelihood with unknown w and unknown precision β. For more information: look at exercise 2.11.

5.7 Summary of lecture 3

In this lecture we have introduced the Bayesian linear regression and:

- How to learn the model
- How to use the model for making predictions
- How to use features

We also presented two Bayesian methods for selecting hyperparameters in a probabilistic model by:

- Maximizing marginal likelihood
- Using hyperpriors