Advanced Probabilistic Machine Learning
SSY316

Basics of Probability Theory

Alexandre Graell i Amat
alexandre.graell@chalmers.se
https://sites.google.com/site/agraellamat

October 31, 2023

CHALMERS

Notation

- Vectors: Bold lowercase, e.g., $\boldsymbol{x}$
- Matrices: Bold uppercase, e.g., $\boldsymbol{X}$
- Random variables, vectors, and matrices: Sansserif font, e.g. x, $\boldsymbol{\mathsf{x}}$, and $\boldsymbol{\mathsf{X}}$
- Sets: Caligraphic letters, e.g., $\mathcal{X}$

Discrete random variables

- Probability mass function: $p_{\mathsf{x}}(x) = \Pr(\mathsf{x} = x) = p(x)$, with

$$0 \leq p(x) \leq 1 \qquad \text{and} \qquad \sum_{x \in \mathcal{X}} p(x) = 1$$

- Joint distribution:

$$p_{\mathsf{x},\mathsf{y}}(x, y) = p(x, y)$$

- Conditional distribution

$$p_{\mathsf{x}|\mathsf{y}}(x|y) = p(x|y)$$

Discrete random variables

- Marginal probability:

$$p(x) = \sum_y p(x, y)$$

In general:

$$p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots x_n) = \sum_{x_i} p(x_1, \ldots, x_n)$$

- Bayes' theorem:

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

Continuous random variables

Probability density function: Describes the probability of the value of a continuous random variable x falling within a given interval.

The probability that x falls in an interval $[a, b]$ is

$$p(a \leq \mathsf{x} \leq b) = \int_a^b p(x) dx$$

We have

$$p(x) \geq 0 \qquad \text{and} \qquad \int_{-\infty}^{\infty} p(x) dx = 1$$

- Marginalization of $p(x, y)$ with respect to y:

$$p(x) = \int_y p(x, y) \mathrm{d}y$$

Expectation, variance, and covariance

- Expectation (average value of $f(x)$ under probability distribution $p(x)$):

$$\mathbb{E}_{\mathsf{x}}[f(\mathsf{x})] = \mathbb{E}[f(\mathsf{x})] = \sum_x p(x) f(x) \quad \text{(discrete)}$$

$$\mathbb{E}[f(\mathsf{x})] = \int p(x) f(x) \mathrm{d}x \quad \text{(continuous)}$$

- Sample mean: Given N points drawn from $p(x)$,

$$\mathbb{E}[f(\mathsf{x})] \simeq \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

with

$$\lim_{N \longrightarrow \infty} \frac{1}{N} \sum_{i=1}^{N} f(x_i) = \mathbb{E}[f(\mathsf{x})]$$

Expectation, variance, and covariance

- For expectations of functions of several variables, we will keep the subscript to indicate the variable averaged over, e.g.,

$$\mathbb{E}_{\mathsf{x}}[f(\mathsf{x},\mathsf{y})] \quad \text{or} \quad \mathbb{E}_{\mathsf{x}\sim p(x)}[f(\mathsf{x},\mathsf{y})]$$

- Conditional expectation:

$$\mathbb{E}_{\mathsf{x}\sim p(x|y)}[f(\mathsf{x})|y] = \mathbb{E}_{\mathsf{x}|\mathsf{y}}[f(\mathsf{x})|y] = \sum_x p(x|y)f(x)$$

- Variance (how much variability there is in $f(x)$ around its expected value):

$$\mathsf{Var}[f(\mathsf{x})] = \mathbb{E}\left[\left(f(\mathsf{x}) - \mathbb{E}[f(\mathsf{x})]^2\right)\right]$$

- Variance of x:

$$\mathsf{Var}[\mathsf{x}] = \mathbb{E}\left[\mathsf{x}^2\right] - \mathbb{E}[\mathsf{x}]^2$$

Expectation, variance, and covariance

- Covariance of x and y (the extent to which x and y vary together):

$$\begin{aligned}\mathrm{Cov}[x, y] &= \mathbb{E}_{x,y}\left[(x - \mathbb{E}[x])(y - \mathbb{E}[y])\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

- Covariance of two random vectors:

$$\mathrm{Cov}[\mathbf{x}, \mathbf{y}] = \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])\left(\mathbf{y}^{\mathsf{T}} - \mathbb{E}[\mathbf{y}^{\mathsf{T}}]\right)\right]$$

Probabilities: Frequentist vs Bayesian view

- Frequentist interpretation: relative frequency of occurrence of an outcome after repeating an experiment a large number of times

$$p = \lim_{n \to \infty} \frac{k}{n}$$

- Bayesian interpretation: quantifies the uncertainty of events happening

Probabilistic reasoning: Example

- 90% of people with Kreuzfeld-Jacob (KJ) disease ate hamburgers
- The probability of an indivitual to have KJ is one in 100000

1. Assuming half of the population eat hamburgers, what is the probability that a hamburger eater will have KJ disease?

$$\mathsf{KJ} \equiv \text{Having Kreuzfeld-Jacob disease}$$
$$\mathsf{H} \equiv \text{Eating Hamburger}$$

$$p(\mathsf{KJ} = \text{yes}|\mathsf{H} = \text{yes}) \ ?$$

Probabilistic reasoning

Interpretation (frequentist approach)

Consider a population of 1M people:

- $p(\mathsf{KJ} = \mathsf{yes}) \Rightarrow 1000000 \cdot (1/10000) = 10$
- $p(\mathsf{H} = \mathsf{yes} | \mathsf{KJ} = \mathsf{yes}) \Rightarrow 10 \cdot 0.9 = 9$
- $p(\mathsf{H} = \mathsf{yes}) = 0.5 \Rightarrow 500000$

	H = yes	H = no
KJ = yes	9	1
KJ = no	499991	499 999

Now:

- $p(\mathsf{KJ} = \mathsf{yes} | \mathsf{H} = \mathsf{yes}) \equiv$ proportion of hamburger eaters having KJ:

$$p(\mathsf{KJ} = \mathsf{yes} | \mathsf{H} = \mathsf{yes}) = \frac{9}{9 + 499991} = 1.8 \cdot 10^{-5}$$

But this can we written as

$$\begin{aligned} p(\mathsf{KJ} = \mathsf{yes} | \mathsf{H} = \mathsf{yes}) &= \frac{p(\mathsf{KJ} = \mathsf{yes}, \mathsf{H} = \mathsf{yes})}{p(\mathsf{KJ} = \mathsf{yes}, \mathsf{H} = \mathsf{yes}) + p(\mathsf{KJ} = \mathsf{no}, \mathsf{H} = \mathsf{yes})} \\ &= \frac{p(\mathsf{H} = \mathsf{yes} | \mathsf{KJ} = \mathsf{yes}) p(\mathsf{KJ} = \mathsf{yes})}{P(\mathsf{H} = \mathsf{yes})} \end{aligned}$$

Probabilistic reasoning

Interpretation (Bayesian approach)

x: our hypothesis (e.g. patient has a disease or not)
y: data (e.g., test results or patient symptoms)

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

- $p(x)$: prior belief in the hypothesis before looking at any data
- $p(y|x)$: likelihood of the data if the hypothesis were true
- $p(y)$: marginal likelihood (commonness of the data)
- $p(x|y)$: posterior belief on a hypothesis given the data

Bayesian (probabilistic) modeling

Two types of variables:

- $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$: Observed variables (the data)
- $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_M\}$: Latent variables (we want to learn)

Probabilistic modeling: Treat both observed and latent variables as random variables.

Can model relationship between $\mathcal{D}$ and $\boldsymbol{\theta}$ via $p(\mathcal{D}, \boldsymbol{\theta})$

Usually we will be interested in $p(\boldsymbol{\theta}|\mathcal{D})$.

Bayesian (probablistic) inference

Many inference problems are of the form:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

- $\mathcal{D}$: Observed data
- $\boldsymbol{\theta}$: parameters of some model explaining the data
- $p(\boldsymbol{\theta})$: prior belief of the parameters before collecting any data
- $p(\mathcal{D}|\boldsymbol{\theta})$: likelihood of the data in view of the parameters
- $p(\mathcal{D})$: marginal likelihood
- $p(\boldsymbol{\theta}|\mathcal{D})$: posterior belief of the parameters after observing the data

Bayesian (probablistic) inference

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}$$

- Seeing quantities as functions of $\boldsymbol{\theta}$, $p(\mathcal{D})$ can be viewed as a normalization constant and we can write

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

- Most probable a posteriori (maximum a posteriori (MAP)) setting:

$$\boldsymbol{\theta}^*_{\mathsf{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathcal{D})$$

- If $p(\boldsymbol{\theta})$ is constant, MAP is equivalent to maximum likelihood,

$$\theta^*_{\mathsf{ML}} = \arg\max_{\theta} p(\mathcal{D}|\theta)$$

Example: Tossing a biased coin

- x $\in \{0, 1\}$: Outcome of a coin flip ($0 \equiv$ tail, $1 \equiv$ head)

$$p(\mathsf{x} = 1) = \mu, \qquad p(\mathsf{x} = 0) = 1 - \mu$$

Goal: Given a data set $\mathcal{D} = \{x_1, \ldots, x_N\}$, estimate μ, i.e., the probability that a toss coin will be a head, $p(\mu|\mathcal{D})$.

- Solution: Apply Bayes',

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

Example: Tossing a biased coin

- Observation: A single coin toss corresponds to a Bernoulli RV,

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

with

$$\mathbb{E}[\mathsf{x}] = \mu, \qquad \mathsf{Var}[\mathsf{x}] = \mu(1-\mu)$$

- For N coin tosses,

$$\begin{aligned} p(\mathcal{D}|\mu) &= \prod_{i=1}^{N} p(x_i|\mu) = \prod_{i=1}^{N} \mu^{x_i}(1-\mu)^{1-x_i} \\ &= \mu^{\sum_i x_i}(1-\mu)^{N-\sum_i x_i} = \mu^{h}(1-\mu)^{N-h} \end{aligned}$$

where $h = \sum_{i=1}^{N} x_i$ is the number of heads

Example: Tossing a biased coin

Frequentist approach:

Can estimate μ by maximizing $p(\mathcal{D}|\mu)$ or, equivalently $\ln p(\mathcal{D}|\mu)$,

$$\begin{aligned}\ln p(\mathcal{D}|\mu) &= \sum_{i=1}^{N} \ln p(x_i|\mu) = \sum_{i=1}^{N} \ln\left(\mu^{x_i}(1-\mu)^{1-x_i}\right)\\ &= \sum_{i=1}^{N} x_i \ln\mu + (1-x_i)\ln(1-\mu)\end{aligned}$$

Differentiating and equating to zero we obtain the ML estimator

$$\mu_{\mathsf{ML}} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{h}{N}$$

h: number of heads within data set

Example: Tossing a biased coin

Bayesian approach:

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

with

$$p(\mathcal{D}|\mu) = \mu^h(1-\mu)^{N-h}$$

- Specify a prior for $p(\mu)$!
- Assume $\mu \in \{0.1, 0.5, 0.8\}$ with

$$p(\mu = 0.1) = 0.15, \qquad p(\mu = 0.5) = 0.8, \qquad p(\mu = 0.8) = 0.05$$

$N = 10$ with 2 heads and 8 tails

$$p(\mu = 0.1|\mathcal{D}) = 0.4525 \qquad p(\mu = 0.5|\mathcal{D}) = 0.5475 \qquad p(\mu = 0.8|\mathcal{D}) = 0.00001$$

$N = 100$ with 20 heads and 80 tails

$$p(\mu = 0.1|\mathcal{D}) = 0.99999807 \quad p(\mu = 0.5|\mathcal{D}) = 1.93 \cdot 10^{-6}$$
$$p(\mu = 0.8|\mathcal{D}) = 2.13 \cdot 10^{-35}$$

Example: Tossing a biased coin

And if we consider a continuum of parameters?

A flat (uniform) prior $p(\mu) = k$:

- For continuous variables, we require

$$\int p(\mu)d\mu = 1 \quad \Longrightarrow \quad \int_0^1 p(\mu)d\mu = k = 1$$

- Now:

$$p(\mu|\mathcal{D}) \propto p(\mathcal{D}|\mu)p(\mu) = \mu^h(1-\mu)^{N-h}$$

We want $p(\mu|\mathcal{D})$ to be a distribution,

$$p(\mu|\mathcal{D}) = \frac{1}{c}p(\mathcal{D}|\mu)p(\mu) = \frac{1}{c}\mu^h(1-\mu)^{N-h}$$

where constant c is obtained as

$$c = \int_0^1 \mu^h(1-\mu)^{N-h}d\mu \equiv \mathsf{B}(h+1, N-h+1)$$

The Beta distribution

- Beta function:

$$\mathsf{B}(a,b) = \int_0^1 \mu^{a-1}(1-\mu)^{b-1} dt$$

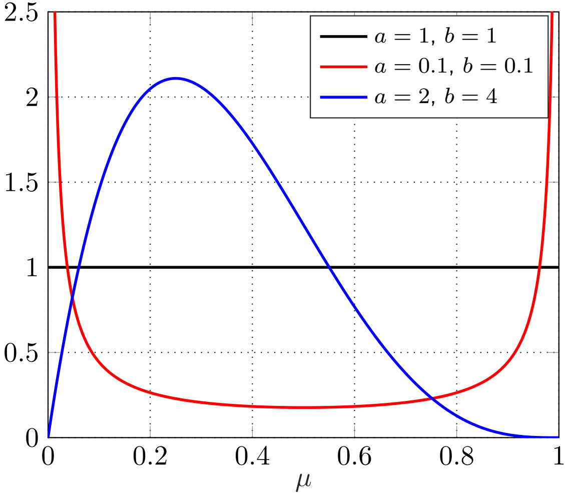
- Beta distribution:

$$\begin{aligned}\mathsf{Beta}(\mu; a, b) &= \frac{1}{\mathsf{B}(a,b)}\mu^{a-1}(1-\mu)^{b-1} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}\end{aligned}$$

$\Gamma(\cdot)$: Gamma function
$\mathsf{B}(a,b)$ A normalization constant to ensure

$$\int_0^1 \mathsf{Beta}(\mu; a, b) d\mu = 1$$

The Beta distribution



- a and b control the distribution of μ (hyperparameters)

Example: Tossing a biased coin

Observation: If prior proportional to powers of μ and $1-\mu$, then posterior distribution will have the same functional form as the prior.

A conjugate prior (posterior will be of same functional form as prior):

- Choose Beta distribution for the prior:

$$p(\mu) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}$$

Then

$$\begin{aligned} p(\mu|\mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu) \\ &\propto \mu^{h}(1-\mu)^{N-h}\mu^{a-1}(1-\mu)^{b-1} \\ &= \mu^{h+a-1}(1-\mu)^{N-h+b-1} \end{aligned}$$

The posterior is also a Beta distribution!

$$p(\mu|\mathcal{D}) = \text{Beta}(\mu; a', b')$$

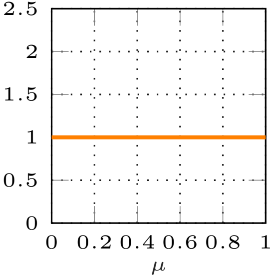
with $a' = a + h$ and $b' = b + N - h$

Example: Tossing a biased coin

prior

$p(\mu) = \text{Beta}(\mu; a, b)$

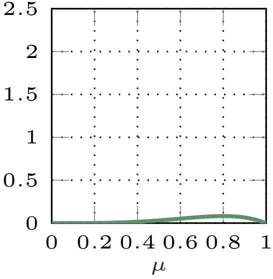
$a = 1,\ b = 1$



likelihood

$p(\mathcal{D}|\mu) = \mu^h (1-\mu)^{N-h}$

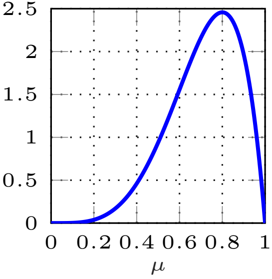
$N = 5,\ h = 4$



posterior

$p(\mu|\mathcal{D}) = \text{Beta}(\mu; a', b')$

$a' = 5,\ b' = 2$



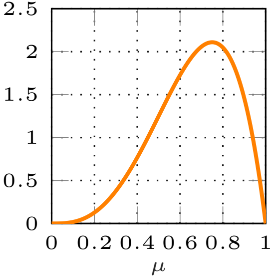
- If we do not know anything about the coin: uniform prior
- $N = 1,\ \mathcal{D} = (1)\ N = 5,\ \mathcal{D} = (1, 1, 1, 0, 1)$

Example: Tossing a biased coin

prior

$p(\mu) = \text{Beta}(\mu; a, b)$

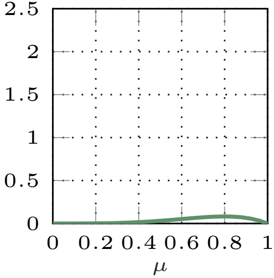
$a = 4,\ b = 2$



likelihood

$p(\mathcal{D}|\mu) = \mu^h (1-\mu)^{N-h}$

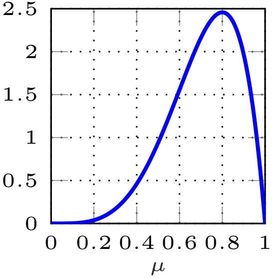
$N = 5,\ h = 4$



posterior

$p(\mu|\mathcal{D}) = \text{Beta}(\mu; a', b')$

$a' = 5,\ b' = 2$



- $N = 5$, $\mathcal{D} = (1, 1, 1, 0, 1)$
- Sequential inference: Posterior can act as prior if we subsequently observe additional data!
- $\mathcal{D} = (1, 1, 1, 0, 1)$

Bayesian inference and machine learning

Probabilistic (Bayesian) machine learning:

- Treats model and its parameters as random variables
- Learning does not provide a single model, but a distribution of likely models
- Can incorporate prior knowledge on model and parameters