Notes on probability theory, Bayes theorem and Bayesian learning

Marta Arias, Computer Science @ UPC

These notes provide a quick reference to various aspects of probability theory, estimation, Bayesian learning and related aspects that we are going to use during the course. Make sure that you understand the concepts and you attempt to solve the exercises proposed.

Disclaimer: these are by no means a complete or mathematically super-rigorous. For that, please consult proper books.

You can find some nice introductory videos of MIT's course "Intro to Probability".

1. Probability theory basics

Let Ω be a set of possible outcomes of an event, and $A \subseteq \Omega$ an **event**. A **probability measure** $P : \mathcal{P}^{\Omega} \to \mathbb{R}$ assigns a real number to every subset of Ω; namely $P(A)$ represents how likely it is that the experiment's outcome falls in A.

Axioms

The three axioms of probability are:

- $P(A) \geq 0$ for all events $A \subseteq \Omega$
- $P(\Omega) = 1$
- $P(A \cup B) = P(A) + P(B)$ for **disjoint** events A and B (i.e. $A \cap B = \emptyset$)

Some consequences:

- $P(\overline{A}) \stackrel{\text{def}}{=} P(\Omega \setminus A) = 1 - P(A)$
- $P(\emptyset) = 0$
- If $A \subseteq B$, then $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cup B) \leq P(A) + P(B)$
- and so on

Oftentimes we are interested in measuring how likely two (or more) events happen simultaneously, in this case we say that the **joint** probability of events A and B is $P(A, B) \stackrel{\text{def}}{=} P(A \cap B)$. Namely we use the "comma" as an **and** or **intersection**.

Sum rule

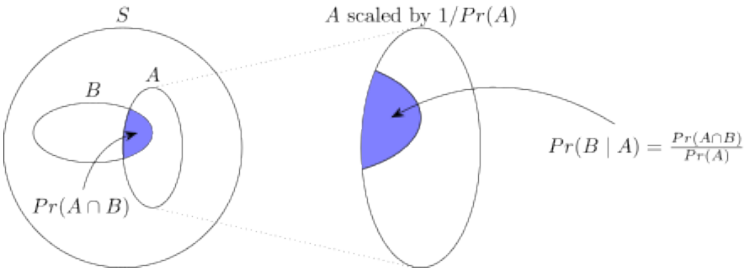
If $A \subseteq B_1 \cup .. \cup B_n$, and $B_i \cap B_j = \emptyset$ for all $1 \leq i < j \leq n$, then:

$$P(A) = \sum_{i=1}^{n} P(A, B_i)$$

This is also known as **marginalization**; when we have a joint probability $p(x, y)$ and want to compute $p(x)$ we way that we "marginalize out" y, which means applying the sum rule $p(x) = \sum_y p(x, y)$ or $p(x) = \int_y p(x, y)dy$ if y is continuous.

Conditional probability

The conditional probability of B given A (provided $P(A) > 0$) is given by $P(B|A) \stackrel{\text{def}}{=} \frac{P(A,B)}{P(A)}$



Notice that rearranging terms leads us to the useful **product rule** which allows us to decompose a joint probability into a product of conditionals $P(A, B) = P(B|A)P(A) = P(A|B)P(B)$.

General product rule

This rule helps us to break down (i.e. *factorize*) a (possibly large) joint distribution into the product of smaller **conditional** distributions.

$$P(A_1, A_2, \ldots, A_n) = \prod_{i=1}^{n} P(A_i|A_1, .., A_{i-1})$$

Exercise 1. Prove that $P(A, B, C) = P(A) \times P(B|A) \times P(C|A, B) = P(C) \times P(B|C) \times P(A|B, C)$. Then, prove the general claim above. You shoud use the simple product rule repeatedly.

You should definitely remember these two rules:

- **Sum rule:** $P(A) = \sum_i P(A, B_i)$
- **Product rule:** $P(A, B) = P(A|B) \times P(B)$

Bayes rule (or theorem)[1]

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

Exercise 2. Prove Bayes rule.[2]

[1] Look at this for a nice visual explanation of Bayes theorem.
[2] Solution here

Example of Bayes rule in action

An English-speaking tourist visits a city whose language is not English. A local friend tells him that 1 in 10 natives speak English, 1 in 5 people in the streets are tourists and that half of the tourists speak English. Our visitor stops someone in the street and finds that this person speaks English. What is the probability that this person is a tourist?

2. Random variables and distributions

A **random variable** is a variable whose value is determined by chance. A **discrete r.v.** takes on a countable number of distinct values, for example the result of tossing a coin or a dice. A **continuous r.b.** takes on any value in some continuous interval (from the real line, e.g.); examples of continuous r.v. are typically measurments such as the height of a person, temperature during the day, etc.

A **distribution** in statistics is a function that shows the possible values that a random variable can take and how often they occur. Each probability distribution is associated with a graph describing the likelihood of occurrence of every event.

Discrete distributions

These are functions describing the distribution governing a discrete random variable. They are also called *probability mass functions* (pmf). Typical examples are: Bernoulli, Binomial, Geometric, Categorical, Multinomial, Poisson, Zeta, etc.

Continuous distributions

These are functions describing the distribution governing a continuous random variable. They are called *probability density functions* (pdf). Typical examples are: Normal (Gaussian), Uniform, Beta, Laplace, Exponential, Student's t-dist, Dirichlet, Gamma, Wishart, etc.

Expectation, variance and covariance

The **expected value** of a function $g : \mathbb{R} \to \mathbb{R}$ of a univariate continuous random variable $X \sim p(x)$ with support $\mathcal{X}$ is given by

$$\mathbb{E}_X[g(x)] = \int_{\mathcal{X}} g(x)p(x)dx$$

If X is discrete, then this becomes

$$\mathbb{E}_X[g(x)] = \sum_{x \in \mathcal{X}} g(x)p(x)$$

As a special case where g is the *identity*, we get the **expectation** of a continuous random variable:

$$\mathbb{E}[x] = \int_{\mathcal{X}} xp(x)dx$$

or in the discrete case:

$$\mathbb{E}[x] = \sum_{x \in \mathcal{X}} xp(x)$$

The **variance** of a random variable X is defined as:

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

The **covariance** of two random variables X and Y is:

$$Cov[x, y] = \mathbb{E}[(x - \mathbb{E}[x])]\mathbb{E}[(y - \mathbb{E}[y])] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

Useful properties:

$$
\begin{aligned}
\mathbb{E}[ax+b] &= a\mathbb{E}[x]+b \\
\mathbb{E}[ax+b] &= a\mathbb{E}[x]+b \\
Var[x+y] &= Var[x]+Var[y]+2Cov[x,y] \\
Var[x-y] &= Var[x]+Var[y]-2Cov[x,y] \\
Var[ax+b] &= a^2 Var[x]
\end{aligned}
$$

Statistical independence

Two random variables X, Y are statistically **independent** if and only if

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

Intuitively, two random variables X and Y are independent if the value of y (once known) does not add any additional information about x (and vice versa). If X, Y are (statistically) independent, then:

$$
\begin{aligned}
p(x|y) &= p(x) \\
p(x,y) &= p(x)p(y) \\
Var[x+y] &= Var[x]+Var[y] \\
Cov[x,y] &= 0
\end{aligned}
$$

2. Bayes rule in the context of learning

Bayes rule allows us to reason about hypotheses (models) from data (observations):

$$\boxed{P(hypothesis|data) = \frac{P(data|hypothesis)P(hypothesis)}{P(data)}}$$

In our language of paramters and datasets this is: let θ be a random variable with support Θ, e.g. describing the parameters of a modelling function and let D be the data that has been observed. Then:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} = \frac{P(D|\theta)P(\theta)}{\int_\Theta P(D|\theta)P(\theta)d\theta}$$

- $P(\theta)$: **prior** distribution of θ, that is, prior to observing D, what values do we think are more plausible
- $P(D|\theta)$: **likelihood** of θ, it is the probability of observing D if parameters are θ. **Not a distribution over** θ.
- $P(D)$: **evidence** or expected likelihood
- $P(\theta|D)$: **posterior**; quantity of interest, expresses what we know about θ after having observed D

The way we are going to use Bayes rule most often is in the context of, given some data D, finding good values for the unknown parameters θ. In what follows, we have two ways in which we can find (estimate) such values for the parameters which are **maximum likelihood** and **maximum a posteriori**.

Maximum likelihood, as the name suggests, uses the value of θ that maximizes $P(D|\theta)$ i.e. the *likelihood* of θ:

$$\theta_{ML} = \arg\max_\theta = P(D|\theta)$$

Maximum a posteriori takes into account a *prior distribution* for θ and estimates its value using:

$$\theta_{MAP} = \arg\max_\theta = P(D|\theta)P(\theta)$$

Using the posterior

Once we know $P(\theta|C)$ we can do a lot of different things of interest, e.g.:

- Compute the **credible interval** (Bayesian CI)
- Compute the MAP (maximum a posteriori)
- Draw observations from θ
- Compute the expected value of the posterior
- Compute the predictive distribution (**predictive posterior**)

3. Maximum likelihood estimation

One of the workhorses in classical statistical inference, the **maximum likelihood estimation** method is a general principle for estimating parameters.

Given a data sample $D = \{x_1, x_2 \ldots, x_n\}$ where each x_i is an *independent and identically distributed* (iid) observation from a random variable $X \sim p(X; \theta)$ with θ being the parameters of the distribution. Since all x_i are independent, the probability of obtaining the sample D can be expressed as:

$$p(D; \theta) = \prod_{i=1}^{n} p(x_i; \theta)$$

The **likelihood function** is defined as $\mathcal{L}(\theta) = p(D; \theta)$. Notice that the likelihood is a *function* of the parameters θ and not a probaability distribution; it assumes the data D is **fixed**. So, the maximum likelihood estimator for parameters θ is given by:

$$\theta_{ML} \stackrel{\text{def}}{=} \arg\max_{\theta} \mathcal{L}(\theta)$$

Notice that the likelihood is defined as a *product* of probabilities, and given that they tend to be small multiplying a large number of them (if we have many observations) this will end up being a tiny number which may lead to underflows when computing them in a machine. So, it is convenient to work with the *logarithm* of the likelihood function, typically taken to be its negative and then maximum likelihood becomes minimum log likelihood:

$$\theta_{ML} \stackrel{\text{def}}{=} \arg\max_{\theta} \mathcal{L}(\theta) = \arg\min_{\theta} -\log \mathcal{L}(\theta)$$

Exercise 3. Compute the MLE for univariate Gaussian distribution. That is, assume you are given iid observations $D = \{x_1, \ldots, x_n\}$ sampled from $X \sim \mathcal{N}(\mu, \sigma^2)$. Your job is to compute the maximum likelihood estimates μ_{ML} and σ_{ML}. For that define the (minus) log likelihood using the Gaussian probability density function, and find minimums by setting partial derivatives to 0 and solving for μ and σ.

Exercise 4. Compute the MLE for a Bernoulli distribution. Now your observations are the results of n coin tosses. Here, you need to compute the parameter p_{ML} of the Bernoulli random variable. Remember that the probability function of a Bernoulli r.v. is given by $p^x(1-p)^{(1-x)}$ for $x \in \{0, 1\}$.

4. Properties of estimators

Statisticians have defined several desirable properties that *good* estimators should have. Estimators are essentially functions of the data sample D, and as such can be seen as random variables that take on different

values for different D. The following properties essentially define statistics of these functions over the space of possible D.

Unbiasedness An estimator $\hat{\theta}$ is said to be *unbiased* if in the long run it takes on the value of the population (true) parameter.

$$Bias_D[\hat{\theta}] = \mathbb{E}_{\mathrm{D}}[\hat{\theta}] - \theta$$

We say that an estimator is **unbiased** if its $Bias$ is 0, i.e. when $Bias_D[\hat{\theta}] = \theta$. Notice that the expectation is taken over all possible choices of samples D of some fixed size.

Variance This essentially tells us how sensitive $\hat{\theta}$ is to variations of input D (on average):

$$Var_D[\hat{\theta}] = \mathbb{E}_{\mathrm{D}}[\ (\hat{\theta} - \mathbb{E}_{\mathrm{D}}[\hat{\theta}])^2\ ] = \mathbb{E}_{\mathrm{D}}[\hat{\theta}^2] - \mathbb{E}_{\mathrm{D}}[\hat{\theta}]^2$$

We say that an estimate $\hat{\theta}_1$ is **more efficient** than another estimate $\hat{\theta}_2$ if $Var[\hat{\theta}_1] < Var[\hat{\theta}_2]$. Efficiency is obviously a good thing; we want estimators to have small variance (i.e. *robustness is desirable*).

Cramér-Rao bound gives us a theoretical lower bound on the variance of estimates.

Efficiency An estimator is said to be *efficient* if in the class of unbiased estimators it has minimum variance

Consistency A sequence of estimators $\hat{\theta}_n$ is said to be *consistent* if it converges to the true value of the parameter:

$$\forall \epsilon > 0 : \lim_{n \to \infty} Prob(|\theta - \hat{\theta}_n| < \epsilon) = 1$$

If the bias and the variance of an estimator tend to 0 when the size (n) tends to ∞, then it is consistent.

Mean squared error (MSE) The mean squared error of an estimator is defined as:

$$MSE(\hat{\theta}) \stackrel{\text{def}}{=} \mathbb{E}_{\mathrm{D}}[(\theta - \hat{\theta})^2]$$

Exercise 4. Show that $MSE(\hat{\theta}) = Bias_D[\hat{\theta}]^2 + Var_D[\hat{\theta}]$.

Exercise 5. Compute the bias and the variance of the ML estimates μ, σ of a univariate Gaussian. Show that σ_{ML} is biased; we can correct its biasedness by using a different estimator for $\tilde{\sigma}^2 = \frac{n}{n-1}\sigma_{ML}$. Compute the bias and the variance of this new estimator.

5. Maximum a posteriori estimation

Maximum a posteriori estimation generalizes maximum likelihood estimates by taking into account a prior distribution over the parameters to be estimated. In contexts where data is scarce, this is in fact desirable because it regularizes the estimation by imposing a prior belief on how likely each θ is. Thus, **maximum a posteriori** combines evidence from data as given by the likelihood and prior beliefs using Bayes:

$$\hat{\theta}_{MAP} \stackrel{\text{def}}{=} \arg\max_{\theta} P(\theta|D) = \arg\max_{\theta} P(D|\theta)P(\theta)$$

Essentially we can drop the denominator $P(D)$ from the application of Bayes' rule since it does not depend on θ so in our maximization context this is a constant and can be ignored.

Exercise 6: Find the MAP estimate for μ of a univariate Gaussian $X \sim \mathcal{N}(\mu, \sigma^2)$ with Gaussian prior distribution for $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Assume that σ, σ_0, μ_0 are known.

Exercise 7: The maximum likelihood estimate of p for a Bernoulli r.v. X (with possible outcomes being 0 or 1) is given by $\hat{p}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i$ where each $x_i \in \{0, 1\}$ are sampled according to $X \sim \mathcal{B}(p)$. If we have $K > 2$ outcomes - e.g. a die with 6 sides – then we have the categorical distribution also known as *multinoulli* or *generalized Bernoulli* which has support $\{1, \dots, K\}$. Its parameters are $\mathbf{p} = (p_1, \dots, p_K)$ representing the probability of observing each of the possible outcomes (clearly, $0 \leq p_k \leq 1$ for all k, and $\sum_{k=1}^{K} p_k = 1$).

It is convenient to use the *one-of-K encoding*[3] for each outcome. So, the pmf of this distribution becomes $p(\mathbf{x}) = \prod_{i=1}^{K} p_i^{x_i}$. Now, given a sample $D = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of possible outcomes for a multinoulli r.v. X, the maximum likelihood estimate for $\mathbf{p}$ is $\hat{p}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$ for each $k \in \{1, \dots, K\}$. In the previous notation, x_{ik} is the $k-$th entry of $\mathbf{x}_i$, that is, $x_{ik} = 1$ if $\mathbf{x}_i$ corresponds to outcome k and 0 otherwise. We can write this more compactly as $\hat{\mathbf{p}}_{ML} = \frac{1}{n}\sum_i \mathbf{x}_i$.

If some category k is not present in our sample, then its corresponding ML estimate is going to be 0. These 0-estimates are problematic in predictive applications: just because we have not seen a "head" in our sample, it is a very strong statement to say that they are impossible to obtain in further trials. To avoid that, *pseudocounts* are used instead. Pseudocounts represent prior knowledge in the form of (imagined) counts c_k for each category k. The idea is that we assume that the data is *augmented* with our pseudocounts, and then we estimate using maximum likelihood over the **augmented** data, namely: $\hat{\mathbf{p}} = \frac{\mathbf{c}+\sum_i \mathbf{x}_i}{n+\sum_{k'} c_{k'}}$. So, the k'th parameter is estimated as $\hat{p}_k = \frac{c_k+\sum_i x_{ik}}{n+\sum_{k'} c_{k'}}$.

As an example (with 6-sided die, not using one-of-6 encoding but using the digit instead), imagine that we obtain a sample from a die: $\{1, 3, 5, 4, 4, 6\}$. If the vector of pseudocounts is $\mathbf{c} = (1, 1, 1, 1, 1, 1)$, then the estimate $\hat{p}_1 = \frac{\text{"nr. of 1"}+1}{|D|+6} = 1/6$ and $\hat{p}_2 = \frac{\text{"nr. of 2s"}+1}{|D|+6} = 1/12$. Notice that although 2 has not been observed in D, its probability estimate is not 0 but a small number. This special case where all pseudocounts are 1 is known as *Laplace smoothing*.

Prove that using maximum likelihood with pseudocounts corresponds to a MAP estimate with Dirichlet prior with parameters $(c_1 + 1, \dots, c_K + 1)$.

6. Bayesian learning

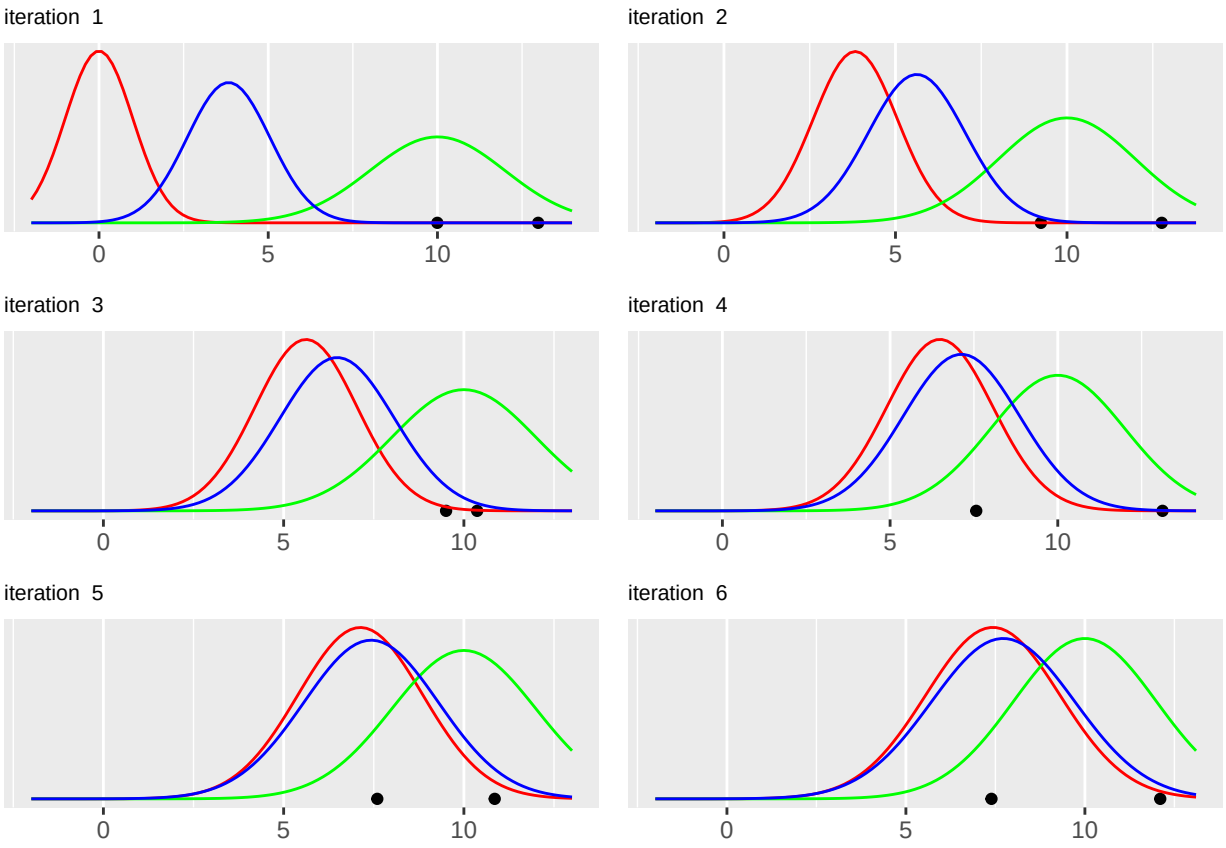
Finally, a glimpse into Bayesian learning. Within this framework, instead of working with point-estimates of our unkown parameter variables θ, we work with the whole (posterior) distribution $P(\theta|D)$.

So we view **learning** as the process by which starting with some prior belief about the parameters (i.e. prior $P(\theta)$) and when facing some observations in the form of a dataset D, we **update our belief** about the possible values for our parameters in the form of the posterior distribution $P(\theta|D)$.

If more data was then given to us, we would continue the process and further update our beliefs about the parameters. Thus, this could be iterated each time we are given new data, and this can be viewed as a sequential process, in which we invoke Bayes each time we need to update our beliefs to obtain a new updated posterior $P(\theta|D_1, D_2, \dots)$.

In the following figure you can see a sequence of updates for a Gaussian univariate random variable. In each plot, the *red* curve shows the prior in red, the true function in green, and the posterior in blue; the data is given by the two dots sampled from the true Gaussian. The next plot in the series has the last posterior as the new prior, and the process follows sequentially. Notice that with more iterations the posterior approaches the true Gaussian.

[3] For example if $K = 3$ then using the one-of-three encoding we would use the $(1, 0, 0)$ vector to encode 1, $(0, 1, 0)$ to encode 2, and $(0, 0, 1)$ to encode 3.



Predictive posterior

Within bayesian learning, when doing predictions the whole distribution $P(\theta|D)$ is used. Namely, we view a prediction as a (weighted) average of all predictions each value for θ can make, weighted by its posterior probability (technically its expected value). Contrast this with other frameworks where just a single point-estimate is used for making predictions (e.g. maximum likelihood or maximum a posteriori):

$$P(x'|D) = \mathbb{E}_{\theta|D}[x] = \int_{\Theta} p(x'|\theta, D)P(\theta|D)d\theta$$

Notice this is a direct application of the product and sum rules.