Probability refresher

Another set of fundamental mathematical tools required to develop various machine learning algorithms (especially towards the end of the course when we will focus on generative modelling)

In order to develop various machine learning algorithms (especially towards the end of the course when we will focus on generative modelling) we need to be familiarized with some basic concepts of: mathematical tools from:

- **Probability**: mathematical framework to handle uncertain statements;
- **Information Theory**: scientific field focused on the quantification of amount of uncertainty in a probability distribution.

Probability

Random Variable: a variable whose value is unknown, all we know is that it can take on different values with a given probability. It is generally defined by an uppercase letter X, whilst the values it can take are in lowercase letter x. (Note: Actually, random variable is not really a variable. To be exact, random variable is actually a function that maps from sample space to the probability space.)

Probability distribution: description of how likely a variable x is, $P(x)$ (or $p(x)$). Depending on the type of variable we have:

- *Discrete distributions*: $P(X)$ called Probability Mass Function (PMF) and X can take on a discrete number of states N. A classical example is represented by a coin where N=2 and $X = 0, 1$. For a fair coin, $P(X = 0) = 0.5$ and $P(X = 1) = 0.5$.

- *Continuous distributions*: $p(X)$ called Probability Density Function (PDF) and X can take on any value from a continuous space (e.g., $\mathbb{R}$). A classical example is represented by the gaussian distribution where $x \in (-\infty, \infty)$.

A probability distribution must satisfy the following conditions:

- each of the possible states must have probability bounded between 0 (no occurrance) and 1 (certainty of occurcence): $\forall x \in X,\ 0 \leq P(x) \leq 1$ (or $p(x) \geq 0$, where the upper bound is removed because of the fact that the integration step δx in the second condition can be smaller than 1: $p(X = x)\delta x <= 1$);

- the sum of the probabilities of all possible states must equal to 1: $\sum_x P(X = x) = 1$ (or $\int p(X = x)dx = 1$).

Joint and Marginal Probabilities: assuming we have a probability distribution acting over a set of variables (e.g., X and Y) we can define

- *Joint distribution*: $P(X = x, Y = y)$ (or $p(X = x, Y = y)$);

- *Marginal distribution*: $P(X=x)=\sum_{y \in Y} P(X=x, Y=y)$ (or $p(X=x)=\int P(X=x, Y=y)dy$), which is the probability spanning one or a subset of the original variables;

Conditional Probability: provides us with the probability of an event given the knowledge that another event has already occurred

$$P(Y=y|X=x)=\frac{P(X=x, Y=y)}{P(X=x)}$$

This formula can be used recursively to define the joint probability of N variables as product of conditional probabilities (so-called *Chain Rule of Probability*)

$$P(x_1, x_2, ..., x_N)=P(x_1)\prod_{i=2}^{N} P(x_i|x_1, x_2, x_{i-1})$$

Independence and Conditional Independence: Two variables X and Y are said to be independent if

$$P(X=x, Y=y)=P(X=x)P(Y=y)$$

If both variables are conditioned on a third variable Z (i.e., P(X=x, Y=y | Z=z)), they are said to be conditionally independent if

$$P(X=x, Y=y|Z=z)=P(X=x|Z=z)P(Y=y|Z=z)$$

Bayes Rule: probabilistic way to update our knowledge of a certain phenomenon (called prior) based on a new piece of evidence (called likelihood):

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

where $P(y)=\sum_x P(x,y)=\sum_x P(y|x)P(x)$ is called the evidence. In practice, it is infeasible to compute this quantity as it would require evaluating y for all possible combination of x (we will see later how it is possible to devise methods for which $P(y)$ can be ignored).

Mean (or Expectation): Given a function $f(x)$ where x is a random variable with probability $P(x)$, its average or mean value is defined as follows for the discrete case:

$$\mu=E_{x\sim P}[f(x)]=\sum_x P(x)f(x)$$

and for the continuous case

$$\mu = E_{x\sim p}[f(x)] = \int p(x)f(x)dx$$

In most Machine Learning applications, we do not have knowledge of the full distribution to evaluate the mean, rather we have access to N equi-probable samples that we assume are drawn from the underlying distribution. We can approximate the mean via the *Sample Mean*:

$$\mu \approx \sum_i \frac{1}{N} f(x_i)$$

Variance (and Covariance): Given a function $f(x)$ where x is a random variable with probability $P(x)$, it represents a measure of how much the values of the function vary from the mean:

$$\sigma^2 = E_{x\sim p}[(f(x) - \mu)^2]$$

Covariance is the extension of the variance to two or more variables, and it tells how much these variables are related to each other:

$$Cov(f(x), g(y)) = E_{x,y\sim p}[(f(x) - \mu_x)(f(y) - \mu_y)]$$

Here, $Cov \to 0$ indicates no correlation between the variables, $Cov > 0$ denotes positive correlation and $Cov < 0$ denotes negative correlation. It is worth remembering that covariance is linked to correlation via:

$$Corr_{x,y} = \frac{Cov_{x,y}}{\sigma_x \sigma_y}$$

Finally, the covariance of a multidimensional vector $\mathbf{x} \in \mathbb{R}^n$ is defined as:

$$Cov_{i,j} = Cov(x_i, x_j), \qquad Cov_{i,i} = \sigma_i^2$$

Distributions: some of the most used probability distributions in Machine Learning are listed in the following.

1. Bernoulli: single binary variable $x \in \{0, 1\}$ (commonly used to describe the toss of a coin). It is defined as

$$P(x = 1) = \phi,\ P(x = 0) = 1 - \phi,\ \phi \in [0, 1]$$

with probability:

$$P(x) = \phi^x (1-\phi)^{1-x} = \phi x + (1-\phi)(1-x)$$

and momentum equal to:

$$E[x] = 1,\ \sigma^2 = \phi(1-\phi)$$

2. Multinoulli (or categorical): extension of Bernoulli distribution to K different states

$$\mathbf{P} \in [0,1]^{K-1};\ P_k = 1 - \mathbf{1}^T\mathbf{P},\ \mathbf{1}^T\mathbf{P} \leq 1$$

3. Gaussian: most popular choice for continuous random variables (most distributions are close to a normal distribution and the central limit theorem states that any sum of independent variables is approximately normal)

$$x \sim \mathcal{N}(\mu, \sigma^2) \rightarrow p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta(x-\mu)^2}{2}}$$

where the second definition uses the precision $\beta = \frac{1}{\sigma^2} \in (0, \infty)$ to avoid possible division by zero. A third way to parametrize the gaussian probability uses $2\delta = log\sigma^2 \in (-\infty, \infty)$ which has the further benefit to be unbounded and can be easily optimized for during training. which is unbounded (compared to the variance that must be positive)

4. Multivariate Gaussian: extension of Gaussian distribution to a multidimensional vector $\mathbf{x} \in \mathbb{R}^n$

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \rightarrow p(\mathbf{x}) = \sqrt{\frac{1}{(2\pi)^n det\boldsymbol{\Sigma}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} = \sqrt{\frac{det\boldsymbol{\beta}}{(2\pi)^n}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\beta}(\mathbf{x}-\boldsymbol{\mu})}$$

where again $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}$. In ML applications, $\boldsymbol{\beta}$ is generally assumed diagonal (mean-field approximation) or even isotropic ($\boldsymbol{\beta} = \beta\mathbf{I}_n$)

5. Mixture of distributions: any smooth probability density function can be expressed as a weighted sum of simpler distributions

$$P(x) = \sum_i P(c=i)P(x|c=i)$$

where c is a categorical variable with Multinoulli distribution and plays the role of a *latent variable*, a variable that cannot be directly observed but is related to x via the joint distribution:

$$P(x,c) = P(x|c)P(c),\ P(x) = \sum_c P(x|c)P(c)$$

A special case is the so-called *Gaussian Mixture* where each probability $P(x|c = i) \sim \mathcal{N}(\mu_i, \sigma_i^2)$.

Information theory

In Machine Learning, we are sometimes interested to quantify how much information is contained in a signal or how much two signals (or probability distributions) differ from each other.

A large body of literature exists in the context of telecommunications, where it is necessary to study how to transmit signals for a discrete alphabet over a noisy channel. More specifically, a code must be designed so to allow sending the least amount of bits for the most amount of useful information. Extension of such theory to continuous variables is also available and more commonly used in the context of ML systems.

Self-information: a measure of information in such a way that likely events have low information content, less likely events have higher information content and independent events have additive information:

$$I(x) = -log_e P(x)$$

such that for $P(x) \rightarrow 0$ (unlikely event), $I \rightarrow \infty$ and for $P(x) \rightarrow 1$ (likely event), $I \rightarrow 0$.

Shannon entropy: extension of self-information to continuous variables, representing the expected amount of information in an event x drawn from a probability $P:

$$H(x) = E_{x \sim P}[I(x)] = -E_{x \sim P}[log_e P(x)]$$

Kullback-Leibler divergence: extension of entropy to 2 variables with probability P and Q, respectively. It is used to measure their distance

$$D_{KL}(P||Q) = E_{x \sim P}[log \frac{P(x)}{Q(x)}] = E_{x \sim P}[logP(x) - logQ(x)] = E_{x \sim P}[logP(x)] - E_{x \sim P}[logQ(x)]$$

which is $D_{KL}(P||Q) = 0$ only when $P = Q$ and grows the further away the two probabilities are. Finally, note that this is not a real distance in that $D_{KL}(P||Q) \neq D_{KL}(Q||P)$ (non-symmetric), therefore the direction matter and it must be chosen wisely when devising optimization schemes with KL divergence in the loss function as we will discuss in more details later.