

Mathematical foundations - probability theory

Andrea Passerini
passerini@disi.unitn.it

Machine Learning

Probability mass function

Given a discrete random variable X taking values in $\mathcal{X} = \{v_1, \dots, v_m\}$, its *probability mass function* $P : \mathcal{X} \rightarrow [0, 1]$ is defined as:

$$P(v_i) = \Pr[X = v_i]$$

and satisfies the following conditions:

- $P(x) \geq 0$
- $\sum_{x \in \mathcal{X}} P(x) = 1$

Discrete random variables

Expected value

- The *expected value*, *mean* or *average* of a random variable x is:

$$E[x] = \mu = \sum_{x \in \mathcal{X}} xP(x) = \sum_{i=1}^m v_i P(v_i)$$

- The *expectation* operator is linear:

$$E[\lambda x + \lambda' y] = \lambda E[x] + \lambda' E[y]$$

Variance

- The *variance* of a random variable is the moment of inertia of its probability mass function:

$$\text{Var}[x] = \sigma^2 = E[(x - \mu)^2] = \sum_{x \in \mathcal{X}} (x - \mu)^2 P(x)$$

- The *standard deviation* σ indicates the typical amount of deviation from the mean one should expect for a randomly drawn value for x .

Properties of mean and variance

second moment

$$E[x^2] = \sum_{x \in \mathcal{X}} x^2 P(x)$$

variance in terms of expectation

$$\text{Var}[x] = E[x^2] - E[x]^2$$

variance and scalar multiplication

$$\text{Var}[\lambda x] = \lambda^2 \text{Var}[x]$$

variance of uncorrelated variables

$$\text{Var}[x + y] = \text{Var}[x] + \text{Var}[y]$$

Probability distributions

Bernoulli distribution

- Two possible values (outcomes): 1 (success), 0 (failure).
- Parameters: p probability of success.
- Probability mass function:

$$P(x; p) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

- $E[x] = p$
- $\text{Var}[x] = p(1 - p)$

Example: tossing a coin

- Head (success) and tail (failure) possible outcomes
- p is probability of head

Proof of mean

$$\begin{aligned}E[x] &= \sum_{x \in \mathcal{X}} xP(x) \\&= \sum_{x \in \{0,1\}} xP(x) \\&= 0 \cdot (1 - p) + 1 \cdot p = p\end{aligned}$$

Proof of variance

$$\begin{aligned}\text{Var}[X] &= \sum_{x \in \mathcal{X}} (x - \mu)^2 P(x) \\ &= \sum_{x \in \{0,1\}} (x - p)^2 P(x) \\ &= (0 - p)^2 \cdot (1 - p) + (1 - p)^2 \cdot p \\ &= p^2 \cdot (1 - p) + (1 - p) \cdot (1 - p) \cdot p \\ &= (1 - p) \cdot (p^2 + p - p^2) \\ &= (1 - p) \cdot p\end{aligned}$$

Probability distributions

Binomial distribution

- Probability of a certain number of successes in n independent Bernoulli trials
- Parameters: p probability of success, n number of trials.
- Probability mass function:

$$P(x; p, n) = \binom{n}{x} p^x (1 - p)^{n-x}$$

- $E[x] = np$
- $\text{Var}[x] = np(1 - p)$

Example: tossing a coin

- n number of coin tosses
- probability of obtaining x heads

Pairs of discrete random variables

Probability mass function

Given a pair of discrete random variables X and Y taking values $\mathcal{X} = \{v_1, \dots, v_m\}$ $\mathcal{Y} = \{w_1, \dots, w_n\}$, the *joint probability mass function* is defined as:

$$P(v_i, w_j) = \Pr[X = v_i, Y = w_j]$$

with properties:

- $P(x, y) \geq 0$
- $\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) = 1$

Properties

- Expected value

$$\mu_x = E[x] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} x P(x, y)$$

$$\mu_y = E[y] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} y P(x, y)$$

- Variance

$$\sigma_x^2 = \text{Var}[(x - \mu_x)^2] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x - \mu_x)^2 P(x, y)$$

$$\sigma_y^2 = \text{Var}[(y - \mu_y)^2] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (y - \mu_y)^2 P(x, y)$$

- Covariance

$$\sigma_{xy} = E[(x - \mu_x)(y - \mu_y)] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x - \mu_x)(y - \mu_y) P(x, y)$$

- Correlation coefficient

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Multinomial distribution (one sample)

- Models the probability of a certain outcome for an event with m possible outcomes.
- Parameters: $p_1, \dots, p_m$ probability of each outcome
- Probability mass function:

$$P(x_1, \dots, x_m; p_1, \dots, p_m) = \prod_{i=1}^m p_i^{x_i}$$

- where $x_1, \dots, x_m$ is a vector with $x_i = 1$ for outcome i and $x_j = 0$ for all $j \neq i$.
- $E[x_i] = p_i$
- $\text{Var}[x_i] = p_i(1 - p_i)$
- $\text{Cov}[x_i, x_j] = -p_i p_j$

Multinomial distribution: example

- Tossing a dice with six faces:
 - m is the number of faces
 - p_i is probability of obtaining face i

Multinomial distribution (general case)

- Given n samples of an event with m possible outcomes, models the probability of a certain distribution of outcomes.
- Parameters: $p_1, \dots, p_m$ probability of each outcome, n number of samples.
- Probability mass function (assumes $\sum_{i=1}^m x_i = n$):

$$P(x_1, \dots, x_m; p_1, \dots, p_m, n) = \frac{n!}{\prod_{i=1}^m x_i!} \prod_{i=1}^m p_i^{x_i}$$

- $E[x_i] = np_i$
- $\text{Var}[x_i] = np_i(1 - p_i)$
- $\text{Cov}[x_i, x_j] = -np_i p_j$

Multinomial distribution: example

- Tossing a dice
 - n number of times a dice is tossed
 - x_i number of times face i is obtained
 - p_i probability of obtaining face i

Conditional probabilities

conditional probability probability of x once y is observed

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

statistical independence variables X and Y are statistical independent iff

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

implying:

$$P(x|y) = P(x)$$

$$P(y|x) = P(y)$$

Basic rules

law of total probability The *marginal distribution* of a variable is obtained from a joint distribution summing over all possible values of the other variable (*sum rule*)

$$P(x) = \sum_{y \in \mathcal{Y}} P(x, y) \qquad P(y) = \sum_{x \in \mathcal{X}} P(x, y)$$

product rule conditional probability definition implies that

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x)$$

Bayes' rule

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

Significance

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

- allows to “invert” statistical connections between *effect* (x) and *cause* (y):

$$\textit{posterior} = \frac{\textit{likelihood} \times \textit{prior}}{\textit{evidence}}$$

- evidence can be obtained using the sum rule from likelihood and prior:

$$P(x) = \sum_y P(x, y) = \sum_y P(x|y)P(y)$$

Playing with probabilities

Use rules!

- Basic rules allow to model a certain probability (e.g. cause given effect) given knowledge of some related ones (e.g. likelihood, prior)
- All our manipulations will be applications of the three basic rules
- Basic rules apply to any number of variables:

$$P(y) = \sum_x \sum_z P(x, y, z) \quad (\text{sum rule})$$

Playing with probabilities

Use rules!

- Basic rules allow to model a certain probability (e.g. cause given effect) given knowledge of some related ones (e.g. likelihood, prior)
- All our manipulations will be applications of the three basic rules
- Basic rules apply to any number of variables:

$$\begin{aligned}P(y) &= \sum_x \sum_z P(x, y, z) \quad (\text{sum rule}) \\&= \sum_x \sum_z P(y|x, z)P(x, z) \quad (\text{product rule})\end{aligned}$$

Playing with probabilities

Use rules!

- Basic rules allow to model a certain probability (e.g. cause given effect) given knowledge of some related ones (e.g. likelihood, prior)
- All our manipulations will be applications of the three basic rules
- Basic rules apply to any number of variables:

$$\begin{aligned}P(y) &= \sum_x \sum_z P(x, y, z) \quad (\text{sum rule}) \\&= \sum_x \sum_z P(y|x, z)P(x, z) \quad (\text{product rule}) \\&= \sum_x \sum_z \frac{P(x|y, z)P(y|z)P(x, z)}{P(x|z)} \quad (\text{Bayes rule})\end{aligned}$$

Example

$$P(y|x, z) = \frac{P(x, z|y)P(y)}{P(x, z)} \quad (\text{Bayes rule})$$

Example

$$\begin{aligned}P(y|x, z) &= \frac{P(x, z|y)P(y)}{P(x, z)} && \text{(Bayes rule)} \\&= \frac{P(x, z|y)P(y)}{P(x|z)P(z)} && \text{(product rule)}\end{aligned}$$

Example

$$\begin{aligned}P(y|x, z) &= \frac{P(x, z|y)P(y)}{P(x, z)} && \text{(Bayes rule)} \\&= \frac{P(x, z|y)P(y)}{P(x|z)P(z)} && \text{(product rule)} \\&= \frac{P(x|z, y)P(z|y)P(y)}{P(x|z)P(z)} && \text{(product rule)}\end{aligned}$$

Example

$$\begin{aligned}P(y|x, z) &= \frac{P(x, z|y)P(y)}{P(x, z)} && \text{(Bayes rule)} \\&= \frac{P(x, z|y)P(y)}{P(x|z)P(z)} && \text{(product rule)} \\&= \frac{P(x|z, y)P(z|y)P(y)}{P(x|z)P(z)} && \text{(product rule)} \\&= \frac{P(x|z, y)P(z, y)}{P(x|z)P(z)} && \text{(product rule)}\end{aligned}$$

Example

$$\begin{aligned}P(y|x, z) &= \frac{P(x, z|y)P(y)}{P(x, z)} && \text{(Bayes rule)} \\&= \frac{P(x, z|y)P(y)}{P(x|z)P(z)} && \text{(product rule)} \\&= \frac{P(x|z, y)P(z|y)P(y)}{P(x|z)P(z)} && \text{(product rule)} \\&= \frac{P(x|z, y)P(z, y)}{P(x|z)P(z)} && \text{(product rule)} \\&= \frac{P(x|z, y)P(y|z)P(z)}{P(x|z)P(z)} && \text{(product rule)}\end{aligned}$$

Example

$$\begin{aligned}P(y|x, z) &= \frac{P(x, z|y)P(y)}{P(x, z)} && \text{(Bayes rule)} \\&= \frac{P(x, z|y)P(y)}{P(x|z)P(z)} && \text{(product rule)} \\&= \frac{P(x|z, y)P(z|y)P(y)}{P(x|z)P(z)} && \text{(product rule)} \\&= \frac{P(x|z, y)P(z, y)}{P(x|z)P(z)} && \text{(product rule)} \\&= \frac{P(x|z, y)P(y|z)P(z)}{P(x|z)P(z)} && \text{(product rule)} \\&= \frac{P(x|z, y)P(y|z)}{P(x|z)}\end{aligned}$$

Continuous random variables

Cumulative distribution function

- How to generalize probability mass function to continuous domains?
- Consider probability of *intervals*, e.g.

$$W = (a < X \leq b) \quad A = (X \leq a) \quad B = (X \leq b)$$

- W and A are mutually exclusive, thus:

$$P(B) = P(A) + P(W) \quad P(W) = P(B) - P(A)$$

- We call $F(q) = P(X \leq q)$ the *cumulative distribution function* (cdf) of X (monotonic function)
- The probability of an interval is the difference of two cdf:

$$P(a < X \leq b) = F(b) - F(a)$$

Probability density function

- The derivative of the cdf is called *probability density function* (pdf):

$$p(x) = \frac{d}{dx} F(x)$$

- The cdf can be computed integrating the pdf:

$$F(q) = P(X \leq q) = \int_{-\infty}^q p(x) dx$$

- Properties:

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

Note

- The pdf of a value x can be greater than one, provided the integral is one.
- E.g. let $p(x)$ be a uniform distribution over $[a, b]$:

$$p(x) = \text{Unif}(x; a, b) = \frac{1}{b-a}(a \leq x \leq b)$$

- For $a = 0$ and $b = 1/2$, $p(x) = 2$ for all $x \in [0, 1/2]$ (but the integral is one)

Properties

expected value

$$E[x] = \mu = \int_{-\infty}^{\infty} xp(x)dx$$

variance

$$\text{Var}[x] = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 p(x)dx$$

Note

Definitions and formulas for discrete random variables carry over to continuous random variables with sums replaced by integrals

Probability distributions

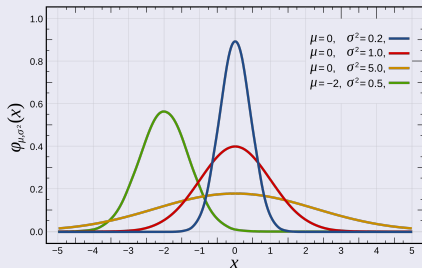
Gaussian (or normal) distribution

- Bell-shaped curve.
- Parameters: μ mean, σ^2 variance.
- Probability density function:

$$p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp - \frac{(x - \mu)^2}{2\sigma^2}$$

- $E[X] = \mu$
- $\text{Var}[X] = \sigma^2$
- Standard normal distribution: $N(0, 1)$
- Standardization of a normal distribution $N(\mu, \sigma^2)$

$$Z = \frac{X - \mu}{\sigma}$$



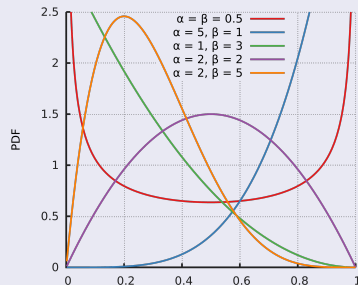
Probability distributions

Beta distribution

- Defined in the interval $[0, 1]$
- Parameters: α, β
- Probability density function:

$$p(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1 - x)^{\beta-1}$$

- $E[x] = \frac{\alpha}{\alpha + \beta}$ $\Gamma(x + 1) = x\Gamma(x), \Gamma(1) = 1$
- $\text{Var}[x] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$



Note

It models the posterior distribution of parameter p of a binomial distribution after observing n independent events with probability p and $1 - p$ with probability $1 - p$.

Probability distributions

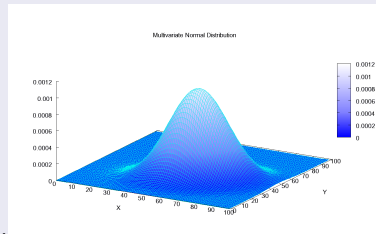
Multivariate normal distribution

- normal distribution for d -dimensional vectorial data.
- Parameters: μ mean vector, Σ covariance matrix.
- Probability density function:

$$p(\mathbf{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp -\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)$$

- $E[x] = \mu$
- $\text{Var}[x] = \Sigma$
- squared *Mahalanobis distance* from $\mathbf{x}$ to μ is standard measure of distance to mean:

$$r^2 = (\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)$$



Probability distributions

Dirichlet distribution

- Defined: $\mathbf{x} \in [0, 1]^m, \sum_{i=1}^m x_i = 1$
- Parameters: $\alpha = \alpha_1, \dots, \alpha_m$
- Probability density function:

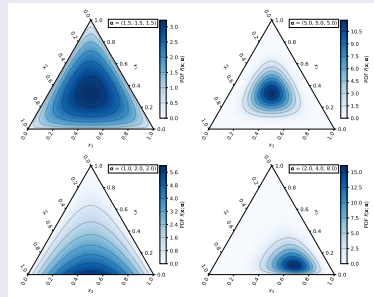
$$p(x_1, \dots, x_m; \alpha) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^m \Gamma(\alpha_i)} \prod_{i=1}^m x_i^{\alpha_i - 1}$$

- $E[x_i] = \frac{\alpha_i}{\alpha_0}$

- $\text{Var}[x_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}$

where $\alpha_0 = \sum_{j=1}^m \alpha_j$

$$\text{Cov}[x_i, x_j] = \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}$$



Note

It models the posterior distribution of parameters $\mathbf{p}$ of a multinomial distribution

Appendix

Additional reference material

Expectation and variance of an average

Consider a sample of $X_1, \dots, X_n$ i.i.d instances drawn from a distribution with mean μ and variance σ^2 .

- Consider the random variable $\bar{X}_n$ measuring the sample average:

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$$

- Its expectation is computed as ($E[a(X + Y)] = a(E[X] + E[Y])$):

$$E[\bar{X}_n] = \frac{1}{n}(E[X_1] + \dots + E[X_n]) = \mu$$

- Its variance is computed as:

$$\text{Var}[\bar{X}_n] = \frac{1}{n^2}(\text{Var}[X_1] + \dots + \text{Var}[X_n]) = \frac{\sigma^2}{n}$$

Expectation of an average

Consider a sample of $X_1, \dots, X_n$ i.i.d instances drawn from a distribution with mean μ and variance σ^2 .

- Consider the random variable $\bar{X}_n$ measuring the sample average:

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$$

- Its expectation is computed as ($E[a(X + Y)] = a(E[X] + E[Y])$):

$$E[\bar{X}_n] = \frac{1}{n}(E[X_1] + \dots + E[X_n]) = \mu$$

- i.e. the expectation of an average is the true mean of the distribution

variance of an average

- Consider the random variable $\bar{X}_n$ measuring the sample average:

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

- Its variance is computed as ($\text{Var}[a(X + Y)] = a^2(\text{Var}[X] + \text{Var}[Y])$ for X and Y independent):

$$\text{Var}[\bar{X}_n] = \frac{1}{n^2}(\text{Var}[X_1] + \cdots + \text{Var}[X_n]) = \frac{\sigma^2}{n}$$

- i.e. the variance of the average *decreases* with the number of observations (the more examples you see, the more likely you are to estimate the correct average)

Chebyshev's inequality

Consider a random variable X with mean μ and variance σ^2 .

- Chebyshev's inequality states that for all $a > 0$:

$$\Pr[|X - \mu| \geq a] \leq \frac{\sigma^2}{a^2}$$

- Replacing $a = k\sigma$ for $k > 0$ we obtain:

$$\Pr[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

Note

Chebyshev's inequality shows that most of the probability mass of a random variable stays within few standard deviations from its mean

Probability laws

The law of large numbers

Consider a sample of $X_1, \dots, X_n$ i.i.d instances drawn from a distribution with mean μ and variance σ^2 .

- For any $\epsilon > 0$, its sample average $\bar{X}_n$ obeys:

$$\lim_{n \rightarrow \infty} \Pr[|\bar{X}_n - \mu| > \epsilon] = 0$$

- It can be shown using Chebyshev's inequality and the facts that $E[\bar{X}_n] = \mu$, $\text{Var}[\bar{X}_n] = \sigma^2/n$:

$$\Pr[|\bar{X}_n - E[\bar{X}_n]| \geq \epsilon] \leq \frac{\sigma^2}{n\epsilon^2}$$

Interpretation

- The accuracy of an empirical statistic increases with the number of samples

Central Limit theorem

Consider a sample of $X_1, \dots, X_n$ i.i.d instances drawn from a distribution with mean μ and variance σ^2 .

- 1 Regardless of the distribution of X_i , for $n \rightarrow \infty$, the distribution of the sample average $\bar{X}_n$ approaches a Normal distribution
- 2 Its mean approaches μ and its variance approaches σ^2/n
- 3 Thus the normalized sample average:

$$Z = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

approaches a standard Normal distribution $N(0, 1)$.

Interpretation

- The sum of a sufficiently large sample of i.i.d. random measurements is approximately normally distributed
- We don't need to know the form of their distribution (it can be arbitrary)
- Justifies the importance of Normal distribution in real world applications

Entropy

- Consider a discrete set of symbols $\mathcal{V} = \{v_1, \dots, v_n\}$ with mutually exclusive probabilities $P(v_i)$.
- We aim at designing a binary code for each symbol, minimizing the average length of messages
- Shannon and Weaver (1949) proved that the optimal code assigns to each symbol v_i a number of bits equal to

$$-\log P(v_i)$$

- The *entropy* of the set of symbols is the expected length of a message encoding a symbol assuming such optimal coding:

$$H[\mathcal{V}] = \mathbb{E}[-\log P(v)] = - \sum_{i=1}^n P(v_i) \log P(v_i)$$

Cross entropy

- Consider two distributions P and Q over variable X
- The *cross entropy* between P and Q measures the expected number of bits needed to code a symbol sampled from P using Q instead

$$H(P; Q) = \mathbb{E}_P[-\log Q(v)] = - \sum_{i=1}^n P(v_i) \log Q(v_i)$$

Note

It is often used as a *loss* for binary classification, with P (empirical) true distribution and Q (empirical) predicted distribution.

Relative entropy

- Consider two distributions P and Q over variable X
- The *relative entropy* or *Kullback-Leibler (KL) divergence* measures the expected length difference when coding instances sampled from P using Q instead:

$$\begin{aligned}D_{KL}(p||q) &= H(P; Q) - H(P) \\&= -\sum_{i=1}^n P(v_i) \log Q(v_i) + \sum_{i=1}^n P(v_i) \log P(v_i) \\&= \sum_{i=1}^n P(v_i) \log \frac{P(v_i)}{Q(v_i)}\end{aligned}$$

Note

The KL-divergence is not a distance (metric) as it is not necessarily symmetric

Conditional entropy

- Consider two variables V, W with (possibly different) distributions P
- The *conditional entropy* is the entropy remaining for variable W once V is known:

$$\begin{aligned} H(W|V) &= \sum_v P(v) H(W|V=v) \\ &= - \sum_v P(v) \sum_w P(w|v) \log P(w|v) \end{aligned}$$

Mutual information

- Consider two variables V , W with (possibly different) distributions P
- The *mutual information* (or *information gain*) is the reduction in entropy for W once V is known:

$$\begin{aligned} I(W; V) &= H(W) - H(W|V) \\ &= - \sum_w p(w) \log p(w) + \sum_v P(v) \sum_w P(w|v) \log P(w|v) \end{aligned}$$

Note

It is used e.g. in selecting the best attribute to use in building a decision tree, where V is the attribute and W is the label.