Mathematical foundations - probability theory

Andrea Passerini
passerini@disi.unitn.it

Machine Learning

Discrete random variables

Probability mass function

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

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

- The *expectation* operator is linear:

$$\mathrm{E}[\lambda x + \lambda' y] = \lambda \mathrm{E}[x] + \lambda' \mathrm{E}[y]$$

Variance

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

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

- 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

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

variance in terms of expectation

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

variance and scalar multiplication

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

variance of uncorrelated variables

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

Probability distributions

Bernoulli distribution

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

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

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

Example: tossing a coin

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

Bernoulli distribution

Proof of mean

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

Bernoulli distribution

Proof of variance

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

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:

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

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

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 $\mathcal{X} = \{v_1, \ldots, v_m\}$ $\mathcal{Y} = \{w_1, \ldots, w_n\}$, 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

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

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

- Variance

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

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

- Covariance

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

- Correlation coefficient

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

Probability distributions

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 $j \neq i$.
- $\mathrm{E}[x_i] = p_i$
- $\mathrm{Var}[x_i] = p_i(1 - p_i)$
- $\mathrm{Cov}[x_i, x_j] = -p_i p_j$

Probability distributions

Multinomial distribution: example

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

Probability distributions

Multinomial distribution (general case)

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

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

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

Probability distributions

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:

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

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

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

Bayes' rule

Significance

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

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

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

- 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 varables:

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

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 varables:

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

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 varables:

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

Playing with probabilities

Example

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

Playing with probabilities

Example

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

Playing with probabilities

Example

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

Playing with probabilities

Example

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

Playing with probabilities

Example

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

Playing with probabilities

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:

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

- 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:

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

Continuous random variables

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

Continuous random variables

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 $[a, b]$:

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

- 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

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

variance

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

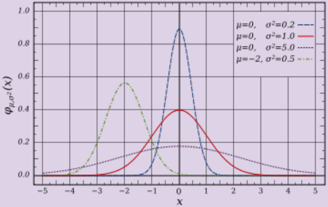
Note

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:



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

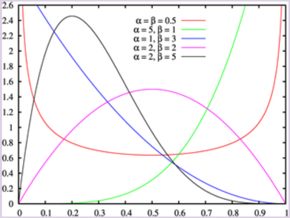
- $\mathrm{E}[x] = \mu$
- $\mathrm{Var}[x] = \sigma^2$
- Standard normal distribution: $N(0, 1)$
- Standardization of a normal distribution $N(\mu, \sigma^2)$

$$z = \frac{x-\mu}{\sigma}$$

Probability distributions

Beta distribution

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



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

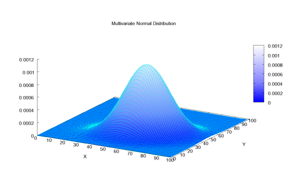
- $\mathrm{E}[x] = \frac{\alpha}{\alpha+\beta}$ $\qquad \Gamma(x + 1) = x\Gamma(x), \Gamma(1) = 1$
- $\mathrm{Var}[x] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Note

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

Probability distributions

Multivariate normal distribution



- normal distribution for d-dimensional vectorial data.
- Parameters: $\boldsymbol{\mu}$ mean vector, Σ covariance matrix.
- Probability density function:

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

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

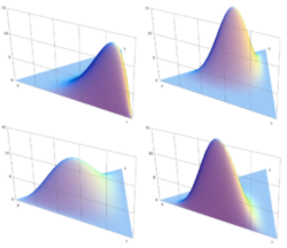
$$r^2 = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

Probability distributions

Dirichlet distribution

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

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



- $\mathrm{E}[x_i] = \frac{\alpha_i}{\alpha_0}$ where $\alpha_0 = \sum_{j=1}^m \alpha_j$
- $\mathrm{Var}[x_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}$ $\mathrm{Cov}[x_i, x_j] = \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}$

Note

It models the posterior distribution of parameters $\boldsymbol{p}$ of a multinomial distribution after observing $\alpha_i - 1$ times each mutually exclusive event

Probability laws

Expectation of an average

Consider a sample of $X_1, \ldots, X_n$ 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:

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

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

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

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

Probability laws

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 $(\mathrm{Var}[a(X+Y)] = a^2(\mathrm{Var}[X] + \mathrm{Var}[Y])$ for X and Y independent):
$$\mathrm{Var}[\bar{X}_n] = \frac{1}{n^2}(\mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n]) = \frac{\sigma^2}{n}$$
- 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)

Probability laws

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 $X_1, \ldots, X_n$ 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:

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

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

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

Interpretation

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

Probability laws

Central Limit theorem

Consider a sample of $X_1, \ldots, X_n$ 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 $N(0, 1)$.

Central Limit theorem

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

Information theory

Entropy

- Consider a discrete set of symbols $\mathcal{V} = \{v_1, \ldots, v_n\}$ with mutually exclusive probabilities $P(v_i)$.
- We aim a 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:

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

Information theory

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

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

Note

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

Information theory

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

Information theory

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:

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

Information theory

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.