A Refresher on Probabilities

Alkis Gotovos, Josip Djolonga

March 1, 2016

Sample spaces and probabilities

- A sample space Ω is the set of outcomes of a random experiment.
- Subsets $A \subseteq \Omega$ are called events.
- For example, consider the experiment of tossing a fair coin twice.
  - Sample space: $\Omega = \{HH, HT, TH, TT\}$
  - Event of at least one "head" occurring: $A = \{HH, HT, TH\}$.
- A probability distribution is a function that assigns a real number $\Pr[A]$ to each event $A \subseteq \Omega$.

Random variables

- Usually, we do not deal directly with sample spaces. Instead, we define random variables and probability distributions on those.
- A random variable is a function $X : \Omega \to \mathbb{R}$.
- For example, if $X :=$ "the number of heads in two coin tosses", then

$$
\begin{aligned}
X(HH) &= 2 \\
X(HT) &= 1 \\
X(TH) &= 1 \\
X(TT) &= 0
\end{aligned}
$$

Probabilities of random variables

- If we denote by $\mathcal{X}$ the set of values a random variable X can take, we can define probabilities directly on $\mathcal{X}$.
- In the above example, $\mathcal{X} = \{0, 1, 2\}$ and we define

$$\begin{aligned}
\Pr[X = 0] &:= \Pr[\{TT\}] \\
\Pr[X = 1] &:= \Pr[\{HT, TH\}] \\
\Pr[X = 2] &:= \Pr[\{HH\}]
\end{aligned}$$

- In practice, we often completely forget about the sample space and work only with random variables.

Discrete random variables

- X is called a discrete random variable if $\mathcal{X}$ is a finite or countably infinite set.
- Examples:
  - $\mathcal{X} = \{0, 1\}$
  - $\mathcal{X} = \mathbb{N}$
  - $\mathcal{X} = \mathbb{N}^d$
- The corresponding probability distribution

$$P(x) := \Pr[X = x]$$

  is called a probability mass function.
- Non-negativity: $P(x) \geq 0, \ \forall x \in \mathcal{X}$
- Normalization: $\sum_{x \in \mathcal{X}} P(x) = 1$

Continuous random variables

- X is called a continuous random variable if $\mathcal{X}$ is an uncountably infinite set.

- Examples:
  - $\mathcal{X} = [0, 1]$
  - $\mathcal{X} = \mathbb{R}$
  - $\mathcal{X} = \mathbb{R}^d$

- The corresponding probability distribution $p(x)$ is called a probability density function.

- Non-negativity: $p(x) \geq 0, \ \forall x \in \mathcal{X}$

- Normalization: $\int_{\mathcal{X}} p(x) dx = 1$

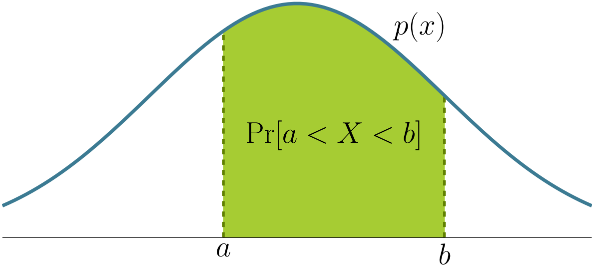
The meaning of density

- Important: For continuous random variables

$$p(x) \neq \Pr[X = x] = 0$$

- To acquire a probability, we have to integrate p over the desired set

$$\Pr[a < X < b] = \int_a^b p(x)dx$$



Joint distributions

- For two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, their joint distribution is defined as

$$P(x, y) := \Pr[X = x, Y = y]$$

- Non-negativity: $P(x, y) \geq 0$
- Normalization: $\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) = 1$
- For example, assume we throw two fair six-sided dice and define $X :=$ "the number on the first die" and $Y :=$ "the number on the second die".
  - $\mathcal{X} = \mathcal{Y} = \{1, 2, 3, 4, 5, 6\}$
  - $P(6, 6) = \Pr[X = 6, Y = 6] = \frac{1}{36}$

Marginal and conditional distributions

Let $P(x, y)$ be a joint distribution of random variables X and Y.

- The marginal distribution of X is defined as

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

- The conditional distribution of X given that Y has a known value y is defined as

$$\begin{aligned} P(x|y) &:= \Pr[X = x | Y = y] \\ &:= \frac{P(x, y)}{P(y)} \qquad \text{(defined if } P(y) > 0\text{)} \end{aligned}$$

- Note that for any fixed y, $P(x|y)$ is a distribution over x, i.e.

$$\sum_{x \in \mathcal{X}} P(x|y) = 1, \ \forall y \in \mathcal{Y}$$

The chain rule

- By definition of conditional distributions, we can always write a joint distribution of X and Y as a product of conditionals:

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

- We can do the same for an arbitrary number of random variables $X_1, \ldots, X_n$:

$$P(x_1, \ldots, x_n) = P(x_1|x_2, \ldots, x_n) \ldots P(x_{n-1}|x_n)P(x_n)$$

Bayes' rule

- For two random variables X and Y, by definition of the conditional distribution of X given Y:

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

- Also, by the chain rule:

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

- Combining the above we get Bayes' rule:

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

Independence

- Two random variables X and Y are called independent, if knowing the value of X does not give any additional information about the distribution of Y (and vice versa):

$$\begin{aligned} & P(x|y) = P(x) \\ \Leftrightarrow \; & P(y|x) = P(y) \end{aligned}$$

- Equivalently, X and Y are independent if their joint distribution factorizes:

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

IID

- IID := Independent and Identically Distributed
- Random variables $X_1, ..., X_n$ are called IID if
  - Each of them has the same (marginal) distribution
  - They are mutually independent
- Note that if $X_1, ..., X_n$ are IID, then

$$\begin{aligned} P(x_1, ..., x_n) &= P(x_1)...P(x_n) \\ &= \prod_{i=1}^{n} P(x_i) \end{aligned}$$

Expectation

- The expectation of a random variable X is defined as

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

- Note that the expectation $\mathrm{E}[X]$ is not the same as the most likely value $\max_{x \in \mathcal{X}} P(x)$.

- Can also be defined for a function f of X:

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

Variance

- The variance of a random variable X is defined as

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

- $\mathrm{Var}[X] \geq 0$

- The standard deviation of X is defined as

$$\sigma_X := \sqrt{\mathrm{Var}[X]}$$

Multidimensional moments

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a vector of random variables.

- The expectation of $\boldsymbol{X}$ is defined as

$$\mathrm{E}[\boldsymbol{X}] := (\mathrm{E}[X_1], \ldots, \mathrm{E}[X_n])$$

- The covariance of variables X_i and X_j is defined as

$$\mathrm{Cov}[X_i, X_j] := \mathrm{E}[(X_i - \mu_{X_i})(X_j - \mu_{X_j})]$$

  - $\mathrm{Cov}[X_i, X_i] = \mathrm{Var}[X_i]$
  - X_i, X_j independent $\Rightarrow \mathrm{Cov}[X_i, X_j] = 0$
  - $\mathrm{Cov}[X_i, X_j] > 0$ roughly means that X_i and X_j increase and decrease together.
  - $\mathrm{Cov}[X_i, X_j] < 0$ roughly means that when X_i increases X_j decreases (and vice versa).

Covariance matrix

For a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ we define its $n \times n$ covariance matrix as follows:

$$\Sigma_{\boldsymbol{X}} = \begin{bmatrix} \mathrm{Var}[X_1] & \mathrm{Cov}[X_1, X_2] & \cdots & \mathrm{Cov}[X_1, X_n] \\ \mathrm{Cov}[X_2, X_1] & \mathrm{Var}[X_2] & \cdots & \mathrm{Cov}[X_2, X_n] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[X_n, X_1] & \mathrm{Cov}[X_n, X_2] & \cdots & \mathrm{Var}[X_n] \end{bmatrix}$$

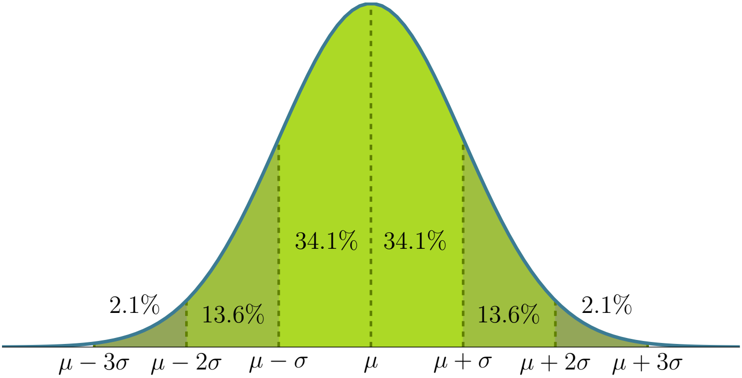
- The diagonal elements are the variances of each random variable $\mathrm{Cov}[X_i, X_i] = \mathrm{Var}[X_i]$.
- $\Sigma_{\boldsymbol{X}}$ is symmetric, because $\mathrm{Cov}[X_i, X_j] = \mathrm{Cov}[X_j, X_i]$.
- $\Sigma_{\boldsymbol{X}}$ is positive semi-definite.
- What does it mean if $\Sigma_{\boldsymbol{X}}$ is diagonal?

Gaussian distribution (1-D)

- Random variable X with $\mathcal{X} = \mathbb{R}$
- Probability density function

$$p(x) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- $\mathrm{E}[X] = \mu,\ \mathrm{Var}[X] = \sigma^2$



Gaussian Distribution (n-D)

- Random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ with $\mathcal{X} = \mathbb{R}^n$
- Probability density function

$$p(\boldsymbol{x}) := \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- $\mathrm{E}[\boldsymbol{X}] = \boldsymbol{\mu}$
- Σ is the covariance matrix of $\boldsymbol{X}$ and $|\Sigma|$ is its determinant.

Data vs. distribution

- Be careful to distinguish between models (usually smooth parametric distributions) and data (sets of points).
- Machine learning:
  - Data = input
  - Distribution = model or assumption
- ML methods usually make some general assumptions about the distribution (e.g. a parametric family), then try to obtain ("infer") the specifics from the data available.
- Example:
  1. Modeling step: Assume a Gaussian distribution as model (parameterized by μ and σ).
  2. Inference step: Estimate parameters μ and σ from data.