Essential Probability

for Machine Learning

Andriy Mnih

Random Variables

- Setup: A hypothetical experiment with an unpredictable outcome is performed.
- A random variable (RV) is a function that assigns a real number to each possible outcome.
- The range $\mathcal{X}$ of a RV X is the set of values it can take.
- A probability mass function $P(X)$ specifies the probability that X is equal to x ($P(X = x)$) for all $x \in \mathcal{X}$.
- $P(X = x)$ is often abbreviated to $P(x)$ when it is clear which RV is referred to.

Why Probability

- Machine learning methods have to deal with uncertainty due to incomplete knowledge and randomness.
- Probability theory provides a framework for dealing with uncertainty.
- Probabilistic methods have been very successful in machine learning. They dominate the field and their popularity keeps growing.

Fundamental Properties of Probability

Suppose X is a discrete RV with range $\mathcal{X}$. Then

- $0 \leq P(X = x) \leq 1$ for all $x \in \mathcal{X}$
- $P(X = x_1$ or $X = x_2) = P(X = x_1) + P(X = x_2)$ for all $x_1, x_2 \in \mathcal{X}$ such that $x_1 \neq x_2$
- $\sum_{x \in \mathcal{X}} P(X = x) = 1$

Discrete and Continuous RVs

- A random variable is discrete if its range is a finite or a countably infinite set. If the range of a RV is uncountable (e.g. $\mathbb{R}$ or $[0, 1]$), the RV is continuous.
- Discrete distributions are specified using probability mass functions, while continuous RVs are specified using probability density functions.
- Most identities that hold for PMFs, hold for PDFs with summation replaced by integration.
- However, there are some differences between PMFs and PDFs:
  - Unlike a PMF, a PDF can take on values greater than 1.
  - The probability of a continuous RV taking on a value in an interval is the area under the PDF in that interval.
  - As a result, the probability of a continuous RV taking on any given value is 0.

Joint Probability

- A joint probability mass function captures dependencies between two or more random variables.
- For two random variables X and Y the joint probability mass function is $P(X = x, Y = y)$, often written as $P(x, y)$.
- The joint PMF has to be normalized so that

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

Marginal Probability

Given the joint distribution of X and Y, the marginal distribution of X can be computed by summing over all the possible values of Y:

$$P(x) = \sum_{y \in \mathcal{Y}} P(x, y). \tag{2}$$

Conditional Probability

- If Y is known to have the value y (and $P(y) \neq 0$), the distribution of X is given by

$$P(x|y) = \frac{P(x, y)}{P(y)}. \tag{3}$$

- Rearranging the definition of conditional probability gives the Chain Rule:

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x). \tag{4}$$

- Note that $P(X|y)$ is a valid probability distribution (just like $P(X)$). As a result,

$$\sum_{x \in \mathcal{X}} P(x|y) = 1, \tag{5}$$

$$0 \leq P(x|y) \leq 1, \text{etc.} \tag{6}$$

Independence

- Random variables X and Y are said to be independent if

$$P(x,y) = P(x)P(y), \tag{7}$$

or, equivalently,

$$P(x|y) = P(x), \tag{8}$$

for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

- Informally, two RVs are independent if knowing the value of one of them provides no information about the value of the other.

Expectation

- The expectation (or expected value or mean) of a random variable X is defined as

$$E[X] = \sum_{x \in \mathcal{X}} xP(x). \tag{10}$$

- Similarly, the expectation of a function $f(X)$ of a random variable X is given by

$$E[f(X)] = \sum_{x \in \mathcal{X}} f(x)P(x). \tag{11}$$

- The expectation of a RV is its “average value", which is not necessarily a “typical" or even valid value for the RV.

Bayes’ Rule

- Bayes’ Rule makes it possible to compute $p(x|y)$ from $p(y|x)$:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} = \frac{P(y|x)P(x)}{\sum_{x \in \mathcal{X}} P(y|x)P(x)} \tag{9}$$

- This is a very important formula, because it can be used to invert “causal" relationships to perform inference.
- A cult was created around this formula in the 18^{th} century.

Properties of Expectation

- Expectation is a linear operator. This means that for $a, b \in \mathbb{R}$

$$E[aX + b] = aE[X] + b, \tag{12}$$

$$E[X + Y] = E[X] + E[Y]. \tag{13}$$

- However, in general, unless X and Y are independent,

$$E[XY] \neq E[X]E[Y]. \tag{14}$$

Variance and Standard Deviation

- The variance of a RV X is defined as

$$Var[X] = E[(X - \mu)^2], \tag{15}$$

  where $\mu = E[X]$.
- Another way to compute $Var[X]$ is

$$Var[X] = E[X^2] - \mu^2. \tag{16}$$

- One important property of variance is $Var[aX + b] = a^2 Var[X]$.
- Variance measures how much a random variable varies about its mean.
- Standard deviation of X is defined as $\sqrt{Var[X]})$.

Energy and Probability

- The energy function corresponding to a probability distribution is defined as

$$E(x) = -\log P(x). \tag{18}$$

- Multiplying probabilities corresponds to adding the corresponding energies.
- Maximizing probability corresponds to minimizing energy.

Covariance

The covariance between X and Y is defined as

$$Cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P(x,y)(x - \mu_X)(y - \mu_Y), \tag{17}$$

where μ_X is the mean of X and μ_Y is the mean of Y.

- If X and Y are independent, $Cov(X,Y) = 0$.
- If the covariance between X and Y is positive and (relatively) large, then X tends to be large when Y is large, and X tends to be small when Y is small.

Bernoulli Distribution

- The PMF of Bernoulli distribution with parameter p is

$$P(x) = \begin{cases} 1 - p & \text{if } x = 0 \\ p & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

- A Bernoulli RV is equal to 1 with probability p and to 0 with probability $1 - p$.
- If X is Bernoulli-distributed,
  - $E[X] = p$
  - $Var[X] = p(1 - p)$

Gaussian Distribution

The PDF of Gaussian distribution with mean μ and variance σ^2 is

$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{20}$$

If X is Gaussian-distributed with the PDF $p(x;\mu,\sigma^2)$ then

- $E[X] = \mu$
- $Var[X] = \sigma^2$

Gaussian (or Normal) distribution is the most widely used continuous distribution.

Entropy

- Entropy measures the amount of uncertainty about the value of a RV.
- The entropy of a RV X (or of a distribution $P(X)$) is given by

$$H(X) = -\sum_{x \in \mathcal{X}} P(x) \log P(x). \tag{21}$$

- The entropy of X does not depend on the values X can take. It depends only on their probabilities.
- The units of entropy are bits if $\log$ is base 2, and nats is $\log$ is base e.

More About Entropy

- Entropy is nonnegative: $H(X) \geq 0$.
- If a RV takes on some value with probability 1, the entropy of that RV is 0, as there is no uncertainty about its value.
- A uniform distribution has the highest entropy of all distributions (over the same set of values).

Kullback-Leibler Divergence

- Kullback-Leibler (KL) divergence between distributions $P(X)$ and $Q(X)$ is defined as

$$KL(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \tag{22}$$

- KL measures the “distance" between distributions.
- Note that KL divergence is not symmetric: in general $KL(P||Q) \neq KL(Q||P)$. As a result, KL divergence is not a true distance metric.
- $KL(P||Q) = 0$ if and only if $P(x) = Q(x)$ for all $x \in \mathcal{X}$.

References

- Introduction to Machine Learning by Ethem Alpaydin
- Review: Probability and Statistics by Sam Roweis