Fundamental concepts

Fundamental concepts regarding Statistics and probability and ML

0.1 Basics

- Probability is the study of uncertainty.
- Probability can be thought of as the fraction of times an event occurs over a large number of experiments or trials.
- We then use this probability as a measure of the chance of some event occurring.
- to begin quantifying uncertainty, we must introduce ourselves to **random variable** which are functions that map the possible set of outcomes of an experiment to the real number space.
- We note that associated with the random variable is a function that measures the probability of the occurrence of an outcome. This function is called the **probability distribution**.
- The **sample space** written as Ω is the set of all possible outcomes of a random experiment. For two coin tosses this would be $\{HT, HH, TH, TT\}$.
- The **event space** is the space of potential results of the experiment. A subset A of Ω is in the event space if at the end of the experiment we get an outcome $\omega \in \Omega$ such that this outcome is in set A. The event space is often the **power set** of the sample space.
- With each event $A \in$ event space we associated a number $P(A)$ that measures the probability or degree of belief that the event will occur.
- The probability of a single event must lie in the interval $[0, 1]$ and the probability of the sample space is 1.
- In the world of machine learning, we use probabilities related to certain quantities of interest denoted by T which can also be said to be our **target space**. The elements of T are called **states**.

- We can define a function such that $X : \Omega \to T$ such that it takes an outcome in the sample space and returns its associated quantity of interest x in T. This mapping from Ω to T is called a **random variable**.
- If a random variable is defined to be the 'number of heads in two coin tosses' then the possible outcomes it can take are: $(0, 1, 2)$. We can see the mapping as: $X(hh) = 2$.
- Basically it is the probabilities of the elements of T that we are actually interested in.

0.2 Discrete

- When the target space is discrete, then the probability that a random variable X takes on a value $x \in T$ is given as $P(X = x)$. This expression is known as the **probability mass function** of random variable X.
- When the target space is continuous then it is more sensible to specify the probability of variable X taking on a value in a certain interval as $P(a \leq X \leq b)$.
- By convention if we denote that probability that X takes on a value less than a specific value x then that function is called a **cumulative distribution function** given by: $P(X \leq x)$.
- Distributions concerned with a single random variable are called **univariate distributions** and distributions of multiple random variables are called **multivariate distributions**.
- Considering discrete random variables, say X and Y. We say that the **target space of the joint probability of these variables is the cartesian product of the target spaces of the individual variables**. Joint probability is then defined as:
$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N} \tag{1}$$
We say that n_{ij} is the number of events with state x_i and y_j and N is the total number of all possible events.
- Joint probability is the probability of the **intersection of two events** that is $P(X = x_i, Y = y_j) = P(X = x_i \cap Y = y_j)$. This represents the joint **probability mass function** and is denoted as $p(x, y)$.
- The **marginal probabiity** that X takes the value x irrespective of the value taken by Y is denoted as $p(x)$ and similarly for Y it is $p(y)$.
- If we only consider instances of $X = x$ then the **conditional probability** or the fraction of instances for which $Y = y$ is denoted as $p(y|x)$.
- Discrete probability distributions are used to model **categorical variables** in machine learning applications.

0.3 Continuous

- The **probability density function** of a continuous random variable has the following property for $f : R^D \rightarrow R$. Where f is the probability density function and R^D is the set of values that continuous random variable takes on.

$$\int_{R^D} f(x)dx = 1 \tag{2}$$

- The distribution of this random variable can be used to compute the probability of its realisations lying in a certain region:

$$P(a \leq X \leq b) - = \int_a^b f(x)dx \tag{3}$$

  Note that the probability of a continuous random variable taking on a particular value $P(X = x)$ is zero.

- When talking about a multivariate random variable, we can think of a set of random variable lying in a vector as $X = [X_1, X_2, \cdots, X_n]^T$ along with its associated vector of realisations as well. Now the **cumulative multivariate distribution** is given by:

$$F_X(x) = P(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_n \leq x_n) \tag{4}$$

  This can be written with multiple integrals as:

$$F_X(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(z_1, z_2, \cdots, z_n)dz_1, \cdots dz_n \tag{5}$$

0.4 Foundations of Bayes

- Before going further note that $p(\boldsymbol{x}, \boldsymbol{y})$ is the joint distribution of two random variables X and Y. $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$ correspond to the marginal distributions and $p(\boldsymbol{y}|\boldsymbol{x})$ is the conditional distribution of y given x.

- The **sum rule** in probability states that:

$$p(\boldsymbol{x}) = \sum_{\forall y} p(\boldsymbol{x}, \boldsymbol{y}) = \int_{\forall y} f(\boldsymbol{x}, \boldsymbol{y})d\boldsymbol{y} \tag{6}$$

  This sum rule basically relates joint distributions to marginal distributions.

- The **product rule** of probability states that:

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y}) \tag{7}$$

  This rule says that every joint distribution can be factorized into a marginal distribution and a conditional distribution.

- Now in machine learning we are often interested in making inferences about unobserved or latent random variables given that we have observed some other random variables. Let us first assume that we have some prior knowledge $p(\boldsymbol{x})$ about the unobserved random variable $\boldsymbol{x}$ and we are also aware of some relationship $p(\boldsymbol{y}|\boldsymbol{x})$ between randon variables $\boldsymbol{x}$ and $\boldsymbol{y}$, which we can observe easily. Then the **Bayes theorem** helps us draw conclusions about unobserved $\boldsymbol{x}$ after we have observed $\boldsymbol{y}$ as:

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})} \tag{8}$$

  The LHS is called the **posterior**, the marginal distribution of the observed variable y is called the **evidence** and the marginal distribution of the unobserved x is called the **prior**. Lastly the conditional distribution of observed y given unobserved x is called the **likelihood**. Note that the evidence can also be written as:

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = E_X[p(\boldsymbol{y}|\boldsymbol{x})] \tag{9}$$

0.5 Summary stats

- We are often interested in summarizing sets of random variables and comparing pairs of random variables. A **statistic** of a random variable is a deterministic function of that random variable. Mean and Variance are important summary statistic measures.

- The **expected value** of a function $g : R \rightarrow R$ of a univariate continuous random variable $X \sim p(x)$ is given by:

$$E_X[g(x)] = \int_X g(x)p(x)dx = \sum_X g(x)p(x) \tag{10}$$

  Similarly the mean of a random variable X is defined as:

$$E_{X_d}(x_d) = \int_{X_d} x_d p(x_d)dx_d = \sum_{X_d} x_d p(X = x_d) \tag{11}$$

- The **covariance** of two random variables is the **expected product of their deviations from their respective means**:

$$Cov_{X,Y}(x, y) = E_{X,Y}[(x - E_X(x))(y - E_Y(y))] \tag{12}$$

- In machine learning we are often dealing with problems that can be modeled as **identically and independently distributed** random variables $X_1, \cdots X_D$. Independence among a large set of random variables means mutual independence among all possible subsets of the random variables. Identically distributed refers to the fact that the random variables have the same distribution.

0.6 ML part 1

- Suppose that we observe some quantitatibve response Y and p different predictors $X_1, \cdots, X_p$. We assume that there is some relationship between Y and $X = (X_1, \cdots, X_p)^T$ that could be written in the general form:

$$Y = f(X) + \epsilon \tag{13}$$

  Now on a general level this function that connects the input variable to the output variable is unknown to us. Our job then is to estimate f based on observed data. We note that **statistical learning** refers to a set of appoaches that help us estimate f. The error term is representative of various other variables that might affect Y but we don't measure them.

- In general there are two types of errors in machine learning: **reducible** and **irreducible** errors. The estimate of modeling function $\hat{f}$ might not be a perfect estimate of f. This induces some error however this error can be reduced using better techniques. We can say the following:

$$E(Y - \hat{Y})^2 = (f(X) - \hat{f}(X))^2 - var(\epsilon) \tag{14}$$

- Our goal in machine learning is to apply a statistical learning method to the training data to estimate an unknown function f.

- When dealing with estimating linear functions of the form of many regression functions like $Y = \beta X + \epsilon$ then we are essentially dealing with **parametric** estimation because in essence, estimating the functional relationship between Y and X boils down to estimating the parameter β since the linear relationship is assumed.

- Now we could have a situation wherein our model fit is far from true values. If this is so we can attempt to fit more **flexible models** that can fit many different functional forms for f. However such flexible models not only require estimating more parameters but also leads to **overfitting** the data which essentially means that the model follows the random noise or errors too closely.

- In machine learning it is often the case that restrictive models are more **interpretable**. For example it is quite easy to understand the relationship between the dependent and independent variables.

0.7 Statistical learning

- Classic examples of **supervised learning methods** are models like Linear regression wherein for each observation of a predictor x_i, there is an associated response measurement y_i. And then we wish to fit a model that relates the response to the predictors.

- Classic examples of **unsupervised learning methods** are models like Kmeans clustering wherein for each observation we might have a vector of measurements x_i but no associated response measurements y_i.
- Problems that involves a quantitative response variable are called **regression problems** and those involving qualitative response variables are called **classification problems**.
- We measure the **quality of fit** of our model by using a measure that tells us how close the predicted response value from our model is to the true value. This measure is commonly termed as the **mean squared error** (MSE) and is given by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x))^2 \tag{15}$$

- Note that what really interests us is to see the accuracy of model predictions when the model is applied to previously unseen data, as opposed to the training data.
- Ideally we choose a learning method that gives us the best or lowest **test MSE**. It can be given as:

$$Avg(y_0 - \hat{f}(x))^2 \tag{16}$$

- We note that a model's **degrees of freedom** which is a measure of its flexibility.
- Note that when a learning method gives us a small training MSE but a large test MSE then we have essentially overfit our data.
- We can show that the expected test MSE for a given new data point x_0 can be decomposed into three constituent values: The variance of $\hat{f}(x_0)$, the squared bias of $\hat{f}(x)$ and the variance of the error term.

$$E(y_0 - \hat{f}(x))^2 = var(\hat{f}(x)) + [bias(\hat{f}(x))]^2 + var(\epsilon) \tag{17}$$

  Where the bias is given by: $p - E(\hat{p})$ for parameter p. The notion of an **average test MSE** implies that we repeatedly estimate our model using a large number of training sets and with each model compute the predicted value of x_0. Is is the average of the difference of predicted vs. actual of this process.
- **Variance** refers to the average amount by which $\hat{f}$ would change if we estimated it using a different training dataset. If a method has high variance, then small changes in the training data can result in very different expressions for $\hat{f}$. Therefore more flexible methods have high variance.
- On the other hand **bias** refers to the error that is introduced by approximation. Generally more flexible methods have a low bias.
- Note that as we increase the flexibility of a model the bias decreases and the variance increases. Finally we note that this relationship between bias, variance and test MSE is called the **bias-variance trade off**.