[1]

Classic machine learning theory relies heavily on several branches of mathematics, notably, probability, calculus and linear algebra for techniques, proofs and algorithms. In this lecture, we will review the important elements of probability theory as they are used in machine learning. We will review the key terminology of probability theory and machine learning and how they relate to each other. We will also introduce important concepts which we will later rely on to examine the operation of the Naïve Bayes classifier.

[2]

Probability has to do with measuring the likelihood of events occurring or not occurring. There are two main ways of measuring the probability of an event. The first is to make predictions about future events based on what has happened in the past, based on data gathered about past events and their outcomes. The second method is to use estimates of future events, usually based on expert knowledge in the problem domain. For this course, we will focus on the first kind, namely using past event data to develop models to help make predictions about the future. This is the dominant approach in machine learning which is why having existing data is so important when embarking on machine learning problems

[3]

Probability theory developed as a branch of mathematics long before data analytics and machine learning came about so you will find different terminology referenced in probability literature. So, let’s start by reviewing some of the most important concepts we will need later in this course. In probability, a “random variable” is a placeholder that can be bound to some value. That value can be categorical such as “true” or “false” or “low”, “medium” or “high”. Or, the value can be continuous such as any positive, real number. For example, consider the random variable X which can take on the all the outcomes of throwing a single die. In this case we say the “domain” of this variable is the set of possible outcomes one-to-six, the faces of the die. The term “experiment” is the act of generating a new outcome. The rolling of a die to generate a new result is an example of an experiment. An “event” is an experiment whose outcome is bound to a value, fixing the value of the random variable. For example, one of the six possible outcomes of throwing a die. The “sample space” of a domain is the set of all possible combinations of values than can be bound to random variables of that domain. For example, suppose we have two dice having random variables X and Y, then the sample space is all of the possible outcomes from throwing the two dice in experiments.

[3]

In machine learning we have different terminology which have equivalent meaning to the concepts in probability theory. Let’s recap some of the important terms. In machine learning, a dataset is collection of labelled examples. Each example, known as an instance, represents one row from the dataset having several column values known as features. These instances are used in training and testing. Labelling means that the features have been categorised. A target feature is a feature for which we want categorise or make some prediction of its value for data examples that we have not trained with nor tested with.

[5]

Comparing probability terminology to machine learning terminology, we can see some direct equivalents between the two theories. A random variable from probability maps to a feature value in machine learning. In the context of machine learning, a feature takes some value associated with an observation. The sample space of a feature is therefore defined as all of the possible values that a feature can take. Just as with probabilistic random variables, features can be categorical or continuous. The equivalent of an experiment is a dataset instance, in other words, a full row of the dataset representing one or more features. An event is therefore some subset of an experiment, that is, one or more feature values from an data instance.

[6]

The reason we consider probability theory in machine learning is that probability is a powerful analysis tool when exploring data and building classification and predictive models. From any given problem domain, we consider the likelihood of some value for any given feature in that domain to be determined by a probability function which denote as P(). The probability of a value being assigned to a feature is the frequency of the observed events with that value relative to the total number of events, typically the total number of instances in the dataset. When a feature is categorical, we say that its function is a probability mass function. When a feature is continuous, its function is known as a probability density function.

[7]

To illustrate these concepts more concretely, let’s consider a simple set of experiments of throwing two unbiased dice. The experiments are labelled A-E and the outcomes for each throw is denoted by an integer representing the result for each die. For example, on the first throw (experiment A), dice number one showed a three and dice number two showed a two. In the coming examples, you can refer back to this dataset to make sense of the results.

[8]

The simplest probability function to consider is known as prior probability. This means that we can calculate the probably of an event using data we already have observed using the relative frequency of the event in the dataset. For example, the probability that dice number one showed a three in an experiment is simply the number of times we observed this outcome divided by the total number of observations. In our example, this is two fifths or 0.4. This is an example of an event having a single random variable

[9]

Joint probability as a simple extension wherein we can consider events with two or more random variables. This is still a prior probability problem in that, again, we used the relative frequency of the observed events matching the query. For example, considering both dice, the probability of dice number one showing a 4 and dice number two also showing a four is one fifth or 0.2.

[10]

Sometimes we need to calculate the probability of an event given we already know outcome of some other event. We call this conditional probability. For example, the probability that dice number one shows a six given that we have already seen that dice two shows a four is calculated as the number of times that this exact event occurs divided by the number of times we observe dice number two showing a four. This is one half or 0.5. For two random variables X and Y, the conditional probability is the joint probability divided by the prior probability of the given event. For more random variables, this formulation can be extended using the chain rule as we shall see shortly.

[11]

Following directly from the definition of conditional probability, we can derive a rule for joint probability by multiplying across the terms. For two random variables, X and Y the joint probability can be calculated as the product of the conditional probability and the prior probability. Note that the order of these calculations is not significant.

[12]

The chain rule is an extension of the product rule for calculation joint probabilities for multiple random variables in terms of their conditional probabilities. As you can see, this becomes computationally intensive as more random variables need to be considered. The chain rule plays an important role in the formulation of Bayesian classification but, luckily, we will be able to exploit an important optimisation, a subject which we will explore in a future lecture.

[13]

In summary, we have discussed how probability theory is related to machine learning both in terms of their equivalent concepts and the role that probability functions play in the determination of feature values in machine learning datasets. We can exploit the formulations of prior, joint and conditional probabilities to explore datasets of features and, as we shall see, we will be able to use these tools to build simple, yet powerful classifiers and predictive models later in this course.