CmpE 343 Lecture Notes
1: Introduction

Ethem Alpaydın

September 28, 2014

There is randomness in life. Though we have some control over certain things, they are also affected by factors beyond our control and we view the outcomes as happening partly by chance. Whether it rains tomorrow is random, whether I will have a car accident next year is random, whether you will win the lottery is random, and so on.

Though there is randomness in life, nobody likes to leave things to chance, because chance is uncertainty. We all desire to have some control over chance to minimize the effect of uncertainty. That is why for example people buy insurance. There is a chance that you will have a car accident and repairing or replacing the car costs a large amount of money; when you buy insurance, you are certain that you lose some money (you pay for the insurance) but if there is an accident, you do not lose a much larger amount. Insurance eliminates the possible harmful effect of chance and gives you certainty.

There are many things that happen by chance but some of the outcomes may be more likely than others. For example tomorrow's being a rainy day is a more likely outcome in winter than in summer. When we talk about the *probability of an event*, we talk about this, namely, how likely it is for an event to occur. For example, if we have a fair coin and we toss it, the probability that the outcome is heads is equal to the probability that the outcome is tails. That is, if we keep on tossing that same coin, we should see roughly the same number of heads and tails.

We have a *random process* where the outcome cannot be predicted in advance, but we know the set of possible outcomes. For example, we toss a coin and though we do not know what the outcome is, it will be heads or tails. The random process has parameters and the outcome depends on them. Though we are not able to predict what the outcome will be, we can calculate how likely each outcome is using *probability theory*. Then if we want to make a prediction about the future, we can choose the most likely outcome.

Coin tossing is an example where the outcomes are discrete. In certain cases, outcomes are continuous. For example, let us say Ali is a jogger and he runs 10K every weekend and he records his time using his smartphone. Every time he runs, depending on his fitness, the weather, road conditions and so on, his time may be a little bit different. There are many factors that have an affect on his performance so we cannot write his time as a deterministic function but rather as a *probability distribution*. For any particular future run, we are not able to predict what his time will be but using the distribution, we can for example calculate the probability that his time will be, say less than t minutes in any particular run. Or we can say something about his average performance, how much in any particular day he can deviate from the average, or for example, we can predict an interval in which his time is going to lie with a certain confidence.

Statistics goes in the opposite direction of probability theory, namely to infer the parameters from the observed outcomes. Let us say somebody gives us a coin and asks us to test whether it is a fair coin. What we do is we toss the coin a number of times and record a *sample* of observations, which is set of heads and tails, and we check if this set of observations is compatible with the claim that it is a fair coin—If I toss the coin 100 times and observe 55 heads and 45 tails, the coin can be fair, but if I observe 80 heads and 20 tails, I should start having doubt.

There are many different application areas of probability and statistics in all sorts of areas, from biology to astronomy, customer relationship management, search engines, pattern recognition, natural language processing, to name a few. Nowadays with the widespread use of computers and networking, in all fields of life, we can collect large amount of data—this is the age of "big data"—and the aim is to infer useful knowledge from this data. This is known as *data mining*.

One popular application of data mining is *basket analysis*. Let us say we work for a supermarket chain that owns many shops, brick and mortar or virtual, selling thousands of goods to millions of customers. It would be very helpful for the chain to predict what are the goods that will be sold because it will allow the chain to better stock its shops; it will also increase customer satisfaction because they are sure to find the items they need. We believe such a prediction is (at least partially) possible because people do not shop completely at random: If the customer is a parent with a baby, he/she buys a certain subset, for example, milk and diapers, and this subset is different from the subset of a single college student who frequently gives parties at his/her home. People buy more ice cream in summer than in winter. There are patterns/regularities/trends in the data and we hope to predict them by collecting a large sample of customer transactions and analyzing it. One type of information we hope to extract is an *association rule* of the form $X \to Y$, which denotes "Customers who buy item X are also likely to buy item Y." This is basically a problem of *estimation*. As we will see later, this can easily be done by estimating the conditional probabilities. Once we find such rules, we can use them to offer products to customers—this is called *cross-selling*. In such an application, the implementation constraints are also important if we have big data with millions of items and billions of transactions.

Let us consider another application area of statistics, namely *hypothesis testing*. Let us say for a course we have three sections for students from three departments and we would like to test if there is a significant difference between the performance, for example quantified as pass/fail ratio, in these three sections. If the claim that there is no difference holds, there should be more or less an equal distribution of ratios in the three sections; if this is not so, we need to reject the claim. Of course these ratios will almost never be exactly the same, there will always be small deviations, but how much deviation is tolerable and how much is not, is one of the things we are going to discuss in later chapters. Note that for such a comparison to make sense, all the other factors that have an effect on performance, namely instructor, TA, textbook, class conditions, and so on, should be the same in all three sections so that any difference is due to the students, which is the factor we are testing.

The third application area of statistics we are going to discuss is *regression*. Let us say we are given the task of estimating the price of a used car. This is again not a deterministic problem, because there are many factors that affect the price, such as engine power, seating capacity, and so on, but there are also factors are not observable, for example, the degree of honesty of the seller. In regression, the aim is to be able to come up with a function that takes the observable factors as argument and returns the price despite the uncertainty introduced by the unobservable factors. Again, what we will predict will not be an exact value but a probability distribution of price.