Saeideh Khadem - ID: 168145

Monday, February 1, 2021 4:36 PM

Please read the slides of Lecture 2 and carry out the reading assignment on the slides. Please write one page a summary for what you have read.

Summary of the slides.

Slides 8-19 : Some Probability rules reviews.

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

$P(C|x)$ ↑ Investigating about the probability of C when x is already has occured

$P(x)$ ← Predictor prior probability

x can be Summation of many factors like $x_1, x_2, \ldots, x_n$, in that case the likelihood of C to occure will be dependant to all $x_1, x_2, \ldots, x_n$ as follows:

$$P(C|X) = P(x_1|C) \times P(x_2|C) \times \ldots \times P(x_n|C) \times P(C)$$

↖ Class prior Probability

Conditional Probability: $P(A|B) = P(A \cap B)/P(B)$ checks about the probability of B when A has happened.

Bayes rule:

$$P(X=x|Y=y) = \frac{P(X=x, Y=y)}{P(Y=y)} = \frac{P(X=x)P(Y=y|X=x)}{\sum_{x'} P(X=x')P(Y=y|X=x')}$$

↑ from series of conditions, the likelihood of x is checked if y has happened

One example of bayes rule.

Assume the probability of having cancer is 0.05 (5% of people have cancer)

Also assume the probability of being smoker is 0.1 (10% of people are smoker)

Let's say among those who have cancer are smoker too: $P(\text{Smoker}|\text{Cancer}) = 0.2$

now, if we want to find the inverse situation and find what percent of smokers have cancer we need to calculate: $P(\text{Cancer}|\text{smoked}) = P(\text{Smoker}/\text{cancer}) \times \frac{P(\text{Cancer})}{P(\text{Smoker})} = 0.1$

Central limit theorm: This theorem states that the sampling distribution of sample mean approaches a normal distribution as the sample size gets large. nomatter what the shape of the population distribution is. This means you take more samples, specially large ones, your graph of sample mean will look like a normal distribution. Example:

Independant or unconditionally independant: $X \perp Y \Leftrightarrow P(X,Y) = P(X)P(Y)$

Conditional Independence: $X \perp Y|Z \Longleftrightarrow P(X,Y|Z) = P(X|Z)P(Y|Z)$

↑ Z has happened

Pages 20-41 Summary

Monday, February 1, 2021 5:19 PM

All I understood from rest of the slides are including as follows: when data are too much information, say stock market, it's so difficult to understand them, and even if we decide to start analizing those information, it's time Consuming and costly. Also, to approach a solution we will encaunter with soo many solutions that may not be promissing. Moreover managing data demands a very complex system. We may have some security gaps too, i.e. some data are noisy and misleading. Some of them can be low quality and inaccurate The other main proble is to figure out how to use the data for meaning - let's say we were able to apply a methode and analogse the big data, how are we going to use it to improve our business or reach faster to the goal. The reasionable Solution for this case is to ture the data to a measurable outcome. That means to use the natural moded of the system, find the algorithm that describes the behaviour of the system and then be able to Predict the future of data. Other factor in dealing with large data is to keep up withe growth in the Data. Assume stock market, many different type of stocks can be added or removed. So our model must be capable of updating itself - with the new data. here the concept of machine learning and consequently a comprehensive math model becames the center of the focus. Our mathematicd predictive model must get updated with the pace of technology. The other very important factor is data integration which consists of taking data from various sources and combining it to create valuable and usable information. There are afew ways to go about integrating data, including the following approches; Consolidation: Combining the data from various sources in one consolidated data store. Propagation: leveraging applications to copy data from one location to another. Federation: Using a virtual database to create a model to match data from different systems and finally Virtualization: Viewing data in one location, but where the data stored separately. The last item to mension about the importance of having Big data analysis skill is lack of skilled workers. while the technological demand is high and artificial intelligence and data analysis tools are innovating swiftly, the lack of the skilleled workers is causing bottleneck for many companies. So it's important to generate user freindly smart analysing tools.