Data Warehouse Systems

Alexander Schlögl

April 6, 2018

Contents

This is **my interpretation** of the lecture slides. I tried to be very verbose and explain everything, all while removing irrelevant parts from the lecture. Using this you should be able to pass the lecture easily. **However, I do not take responsibility for any bad results and will not be blamed from anyone. This was a lot of work and I did it to save others (especially students of following semesters) from having to do this themselves. Use this summary at your own responsibility.** If you have any feedback, feel free to create an issue on the git. I don't promise I will fix anything, but I will try.

1 Probability Theory

1.1 Basics

The following are a few definitions and explanations for concepts needed in this lecture.

Machine Learning

The sole reason for this course is to train students in the basics needed for machine learning. Why this course is held a semester *after* the actual Advanced Machine Learning course I don't know.

Machine learning is the process of generating a classifying function $y(x)$ from a training set $x_1, ..., x_n$. This classifier can then be used to determine the class a new input x' belongs to. The classifier is *learned* from the training set during the *training phase*. It is then usually evaluated on its performance on a *test set*. Machine learning is very useful for finding complex classification functions, as it removes the need of developing a complex algorithm by approximating the underlying probability distribution of X. The performance of the resulting classifier is heaviliy influenced by how closely the traninig set approximates the actual distribution of X. A classifier's ability to classify new input after training is called its ability to *generalize*.

The two main methods of machine learning are *supervised*, where the training data is labelled with their corresponding classes, and *unsupervised*, where the training data is unlabelled and the classifier splits the training set in different clusters.

As data taken from the real-world is usually not split into training and test data, classifiers are often trained using cross validation. With this method the available data is randomly split into training and test data, according to some desired ratio.

Random Variables

A random variable X is sort of a shorthand for the set of results contained in a probability distribution. The actual discrete probabilities are then given using the form $p(X = x_i)$ (remember that for a continuous case $p(X = x_i) = 0$). In his lecture slides Antonio uses functions ($f(x)$) and random variables (X) sort of interchangeably. This is correct from a mathematical standpoint, but might be confusing for students. Just remember that they are essentially different terms for the same thing.

Even worse, I sometimes use X as a random variable, and sometimes as a set. This is also correct, but again might confuse some readers. However, it should be pretty clear from context which I mean, and also there shouldn't be any cases where it makes any tangible difference.

Expectation

Expectation is exactly what the name says: the weighted average value one can expect when drawing from a random source. It is the average of the resutls, weighted by their

probabilities. The formula for this is as follows: $E(f) = \sum_x p(x)f(x)$ for discrete cases and $E(f) = \int p(x)f(x)dx$ for continuous cases. Given a finite number of points drawn from a distribution, the expectation can be approximated using the mean: $E(f) \approx \frac{1}{N} \sum_{n=1}^{N} f(x_n)$.

Covariance

The *variance* of a function measures how "broad" it is (how much variablility of values it has around its mean value). It is calculated according to the formula $var[f] = E[(f(x) - E[f(x)]^2] = E[f(x)^2] - E[f(x)]^2$.

Covariance expresses the extent to which distributions vary *together*. For two distributions it is given as $cov[x, y] = E_{x,y}[xy] - E[x]E[y]$.

The Gaussian Distribution

The Gaussian or *Normal Distribution* is given by the formula

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{1}{2\sigma^2}(x - \mu)^2\right] \tag{1}$$

1.2 Rules of Probabilities

There are some rules you can use when working with probabilities. Call them the "rules of probability algebra" if you will. I will try to explain them using intuitive terms. The one function we need for this is the count function $\#(X)$, which gives the number of elements contained in a set.

Sum Rule

The sum rule can be used for calculating joint probabilities from a set of random samples. Imagine we draw N samples from discrete random distributions X and Y, counting them and writing the counts into a table. The column index i tells us which value X has ($X = x_i$), and the row index j does the same for Y ($Y = y_j$). Contained in the cell n_{ij} is the count of how often $X = x_i, Y = y_j$ occured. This gives us the formula $P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$ for the joint probability of $X = x_i, Y = y_j$.

The *marginal* probability of $X = x_i$ is $\frac{\#(X=x_i)}{N}$. $\#(X = x_i)$ is the total count in column i, which is the sum of all cells in that column. This gives us the sum rule:

$$P(X = x_i) = \frac{\#(X = x_i)}{N} = \frac{1}{N} \sum_j \#(X = x_i, Y = y_j) = \sum_j P(X = x_i, Y = y_j) \tag{2}$$

Product Rule

Sticking with the example from above, we can also find a rule for calculating *conditional probabilities*. $P(Y = y_j)$ is the sum of counts in row j, divided the total sum of the table. The conditional probability $P(Y = y_j | X = x_i)$ is then n_{ij}, divided by the count in column i:

$$P(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i} = \frac{n_{ij}}{\sum_{j'} n_{ij'}} \tag{3}$$

This gives us the following three equations:

$$P(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i} \tag{4}$$

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N} \tag{5}$$

$$P(X = x_i) = \frac{c_i}{N} \tag{6}$$

resulting in the product rule

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \frac{c_i}{N} = P(Y = y_j | X = x_i) P(X = x_i) \tag{7}$$

1.3 Probability Densities

1.4 Conditional Independence

2 Distribution Estimation

2.1 Bayesian Networks

2.2 Markov Random Fields

2.3 Mixture of Gaussians

2.4 Restricted Boltzmann Machines

2.5 Neural Networks