Summary of Lecture 4

We standardize variables by subtracting the mean and dividing by the standard deviation. The new variable can be interpreted as the number of standard deviations away from the mean.

Let X and Y be two random variables. Let

$$Z_X = \frac{X - E[X]}{\sqrt{\text{Var}(X)}} \quad \text{and} \quad Z_Y = \frac{Y - E[Y]}{\sqrt{\text{Var}(Y)}}.$$

The correlation between X and Y random variables (both discrete and continuous) is given as

$$\rho_{XY} = \text{cor}(X, Y) = E[Z_X Z_Y].$$

For discrete random variables, we can compute

$$E[Z_X Z_Y] = \sum_{\text{all values } (z_x, z_y)} z_x z_y P(Z_X = z_x, Z_y = z_y).$$

The correlation always lies between -1 and 1.

The covariance between two random variables X and Y is given by

$$\sigma_{XY} = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y].$$

We also have that

$$\sigma_{XY} = \rho_{XY} \cdot \sigma_X \cdot \sigma_Y.$$

If two variables are independent, then their correlation is equal to zero. However, the converse is not true. That is, if the correlation is equal to zero, the variables can be dependent.

Suppose that a random variable Y is defined as a linear combination of other random variable. Say $Y = c_0 + c_1 X_1 + c_2 X_2$. Then

$$E[Y] = c_0 + c_1 E[X_1] + c_2 E[X_2]$$

and

$$\text{Var}[Y] = c_1^2 \text{Var}[X_1] + c_2^2 \text{Var}[X_2] + 2 c_1 c_2 cov(X_1, X_2).$$

Remember that expectation of the sum is the sum of expectations. However, variance of a sum is not equal to the sum of variances, unless the random variables are independent of each other.

Naive Bayes classifier is a simple statistical tool that allows us to discriminate (predict) between different categorical variables. Recall the homework question where you used linear regression to predict house price based on its size. Usage of a classifier is similar: we would like to use historical data to find a classifier, which we will use to categorize new observations into a set of predetermined categories (usually 2). For example, when a new email arrives, we would like to classify it as spam or not spam; given a brain image of a patient, we would like to identify whether they have a tumor or not; a credit card company would like to predict whether a customer is likely to pay off the balance or not. In general, we have a number of variables $X_1, X_2, \ldots, X_n$ that describe an observation that we would like to put into categories (for simplicity we will talk about two categories, $Y = 0$ and $Y = 1$). Variables $X_1, X_2, \ldots, X_n$ could be words in an email, or pixels in a brain image, or could describe socio-economic status of a new customer. Given variables $X_1, X_2, \ldots, X_n$ our classifier needs to decide if category 1 is more likely than category 0, that is, we would like to compute $P(Y = y \mid X_1, X_2, \ldots, X_n)$ and compare $P(Y = 0 \mid X_1, X_2, \ldots, X_n)$ to $P(Y = 1 \mid X_1, X_2, \ldots, X_n)$. If $P(Y = 1 \mid X_1, X_2, \ldots, X_n)$ is bigger than $P(Y = 0 \mid X_1, X_2, \ldots, X_n)$ then the new observation is more likely to belong to category 1. Using the Bayes theorem, we have that

$$P(Y = y \mid X_1, X_2, \ldots, X_n) = \frac{P(X_1, X_2, \ldots, X_n \mid Y = y) \cdot P(Y = y)}{P(X_1, X_2, \ldots, X_n)}.$$

Since the marginal distribution $P(X_1, X_2, \ldots, X_n)$ is common in both cases when $Y = 1$ and $Y = 0$, we only need to compare $P(X_1, X_2, \ldots, X_n \mid Y = 1) \cdot P(Y = 1)$ to $P(X_1, X_2, \ldots, X_n \mid Y = 0) \cdot P(Y = 0)$ to decide the category of a new observation. In order to compute the conditional distribution $P(X_1, X_2, \ldots, X_n \mid Y = y)$ a naive conditional independence assumption is made. In particular, Naive Bayes assumes that

$$P(X_1, X_2, \ldots, X_n \mid Y = y) = P(X_1 \mid Y = y) \cdot P(X_2 \mid Y = y) \cdot \ldots \cdot P(X_n \mid Y = y).$$

Therefore, a new observation will be put into class 1 if

$$P(Y = 1) \cdot P(X_1 \mid Y = 1) \cdot \ldots \cdot P(X_n \mid Y = 1) > P(Y = 0) \cdot P(X_1 \mid Y = 0) \cdot \ldots \cdot P(X_n \mid Y = 0)$$

and into class 0 otherwise.

Classification may seem similar to clustering we saw in Lecture 2. However, it is fundamentally different. Classification is used to perform a particular task, for example, predict whether a customer is a bargain hunter or loyal to a particular brand, while clustering is used to explore data. One can apply clustering to data at hand to identify groups of similar observations. Once the groups are identified, we can put labels on them, for example, bargain hunter or loyal to a brand. However, notice that before you started clustering data, you did not know which observations corresponded to bargain hunters and which to loyal customers. However, when performing classification, we have information which category an observation belongs to when learning to classify. For example, we know which emails are spam and which are legitimate.