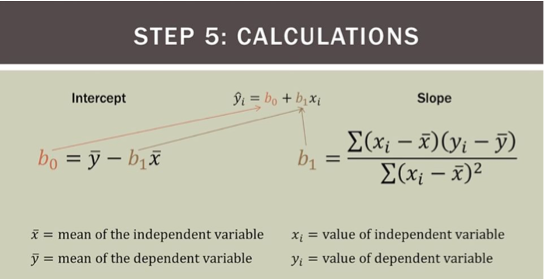
**Data Mining Homework 3 MOHAMMED SHOEBUDDIN HABEEB**

**1)**

**a) Write and explain about the Linear Regression and it’s equation [3 pts]**

**Answer –** Linear Regression is a technique to model a relationship between two features. The result is a linear regression equation that can be used to make predictions of data. Linear regression is a most widely used statistical technique.



*Source: Slides of Data Mining Lecture6-2*

Linear Regression equation is as shown above. yi = b0 + b1xi where b0 is intercept and b1 is slope of the linear equation.

**b) Explain in detail about the loss function of linear regression, R2, Adjusted R2 used in the Linear Regression and what is the need for Adjusted R2? [12 pts]**

**Answer –**

Errors SSR + SSE = SST

Where SSR = Sum of Squares due to Regression,

SSE = Sum of Squares due to Error, and,

SST = Sum of Squares Total

R2 = SSR/SST

Where R2 is the proportion of variation in Y which is explained by the variation in X.

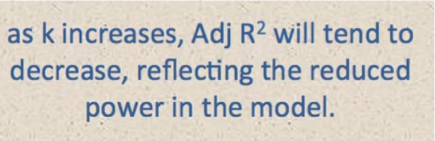
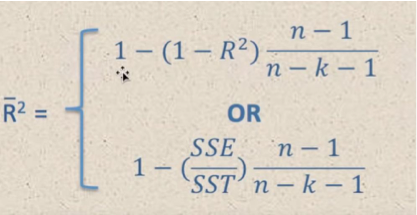
R-squared is the proportion of variance explained. It is the proportion of variance in the observed data that is explained by the model, or the reduction in error over the null model. The null model just predicts the mean of the observed response, and thus it has an intercept and no slope

R-squared is between 0 and 1. Higher values are better because it means that more variance is explained by the model.

For the R-Squared and the Adjusted R-Squared, the closer the value to 1, the better performer our model.

However, R-Square/Adjusted R-Squared doesn’t need to be compared between different models. If the R-Squared/Adjusted R-Squared is .10, we can acknowledge that the model is not doing a great job.

Adjusted R2:-



**c) Please check in** **Shoeb\_Habeeb\_DM\_Homework3.ipynb file.**

**d) Please check in Shoeb\_Habeeb\_DM\_Homework3.ipynb file.**

**e) Please check in Shoeb\_Habeeb\_DM\_Homework3.ipynb file.**

**2)**

**a) What is Conditional probability, Marginal probability and Joint probability? Write their mathematical formulas and give one example each. [5 pts]**

**Answer –**

## ***Conditional Probability:*** A conditional probability is the probability of an event X occurring when a secondary event Y is true. Mathematically, it is represented as P(X | Y). This is read as “probability of X given/conditioned on Y”.

A conditional probability can be calculated as follows:

P(X | Y)=P(X,Y)/ P(Y)

Example:  given that you drew a red card, what’s the probability that it’s a four (p(four|red))=2/26=1/13.  So out of the 26 red cards (given a red card), there are two fours so 2/26=1/13.

### ***Marginal probability:*** The probability of an event occurring (p(A)), it may be thought of as an unconditional probability.  It is not conditioned on another event.  Example:  the probability that a card drawn is red (p(red) = 0.5).  Another example:  the probability that a card drawn is a 4  (p(four)=1/13).

### ***Joint probability:***  P(X and Y).  The probability of event X and event Y occurring.  It is the probability of the intersection of two or more events.  The probability of the intersection of X and Y may be written p(X ∩ Y). Example:  the probability that a card is a four and red =P(four and red) = 2/52=1/26.  (There are two red fours in a deck of 52, the 4 of hearts and the 4 of diamonds).

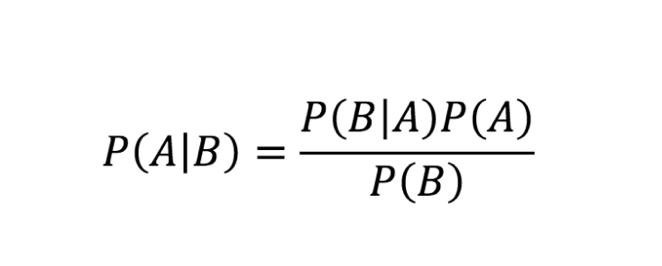
**2) b) Explain what is Baye’s rule with the formula and what is prior, posterior, likelihood and marginal probability in the Baye’s rule. [10 pts]**

**Answer –**

Bayes rule provides us with a way to update our beliefs based on the arrival of new, relevant pieces of evidence.

For example, if we were trying to provide the probability that a given person has cancer, we would initially just say it is whatever percent of the population has cancer. However, given additional evidence such as the fact that the person is a smoker, we can update our probability, since the probability of having cancer is higher given that the person is a smoker. This allows us to utilize prior knowledge to improve our probability estimations.

The below equation is Bayes rule:



The rule has a very simple derivation that directly leads from the relationship between joint and conditional probabilities.

From a Bayesian perspective, we begin with some **prior probability** for some event, and we update this prior probability with new information to obtain a **posterior probability**. The **posterior probability** can then be used as a prior probability in a subsequent analysis.

Bayes’ Theorem, expressed in terms of probability distributions, appears as:

f(θ|data) = f(data|θ)f(θ) / f(data) ,

where f(θ|data) is the posterior distribution for the parameter θ, f(data|θ) is the sampling density for the data—which is proportional to the Likelihood function, only differing by a constant that makes it a proper density function—f(θ) is the prior distribution for the parameter, and f(data) is the **marginal probability** of the data. For a continuous sample space, this **marginal probability** is computed as:

f(data) = Z f(data|θ)f(θ)/dθ,

the integral of the sampling density multiplied by the prior over the sample space for θ.

This quantity is sometimes called the **“marginal likelihood**” for the data and acts as a normalizing constant to make the posterior density. Because this denominator simply scales the posterior density to make it a proper density, and because the sampling density is proportional to the **likelihood function**, Bayes’ Theorem for probability distributions is often stated as:

**Posterior ∝ Likelihood × Prior,**

where the symbol “∝” means “is proportional to.”

**2) c) What is Naive Bayes algorithm and how is related or derived or inspired from Bayes rule? [5 pts]**

**Answer –**

Naïve Bayes is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Below is the formula for calculating the conditional probability.

\textrm{P(H \textbar E) = }  \frac{\textrm{ P(E \textbar H) * P(H)}} {\textrm{P(E)}}

**where**

* P(H) is the probability of hypothesis H being true. This is known as the prior probability.
* P(E) is the probability of the evidence (regardless of the hypothesis).
* P(E|H) is the probability of the evidence given that hypothesis is true.

P(H|E) is the probability of the hypothesis given that the evidence is there.

Naive Bayes classifier assumes that all the features areunrelated to each other. Presence or absence of a feature does not influence the presence or absence of any other feature.

**Pros:**

* It is easy and fast to predict class of test data set. It also performs well in multi class prediction
* When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.
* It performs well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

**Cons:**

* If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.
* On the other side naive Bayes is also known as a bad estimator, so the probability outputs from predict\_proba are not to be taken too seriously.
* Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.