Gmail **Sumit Gupta <montygupta@gmail.com>**

Exam Notes

23 messages

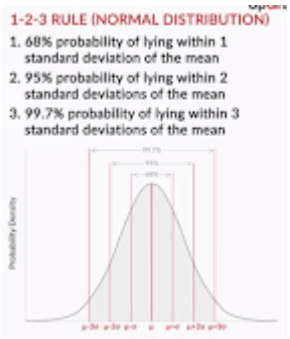
Sumit Gupta <montygupta@gmail.com> Sun, Sep 2, 2018 at 8:07 AM
To: Sumit Gupta <montygupta@gmail.com>

T-distribution:

- A T-distribution is used whenever the standard deviation of the population is unknown
- The degrees of freedom of a T-distribution is equal to sample size n - 1
- For sample size ≥ 30, the T-distribution becomes the same as the normal distribution
- The output values and results of both t-test and z-test are same for sample size ≥ 30

Higher PValue, High Null Value Hypothesis acceptance

---------- Forwarded message ---------
From: **Sumit Gupta** <montygupta@gmail.com>
Date: Sat, Sep 1, 2018 at 5:54 PM
Subject: normal, t, standard, binomial, poisson distribution
To: Sumit Gupta <montygupta@gmail.com>



Screen Shot 2018-09-01 at 5.52.33 PM.png
418K

Sumit Gupta <montygupta@gmail.com> Sun, Sep 2, 2018 at 12:55 PM
To: Sumit Gupta <montygupta@gmail.com>

Types of Derived Metrics:
Type - Nominal, Ordinal (increasing), Interval and Ratio
Business
Data - Like Date
[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Sun, Sep 2, 2018 at 12:58 PM
To: Sumit Gupta <montygupta@gmail.com>

Central Limit Theorem - When you add up multiple independent variables, they show up a normal distribution.
[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Sun, Sep 2, 2018 at 1:04 PM
To: Sumit Gupta <montygupta@gmail.com>

Power Law - Preferential treatment. e.g. Wealth, Names, interest on money etc. money attracts money. These are not normal distributions. Since these are not independent variables, they don't follow Central Limit Theorem.
[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Sun, Sep 2, 2018 at 9:15 PM
To: Sumit Gupta <montygupta@gmail.com>

RSS - Residual Sum of Squares
TSS - Mean Sum of Squares
R2 - 1- TSS/RSS (I if points lie on the line)
RSE = SqRoot RSS/n-2 (2 points will always be straight)
[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Mon, Sep 3, 2018 at 6:29 AM
To: Sumit Gupta <montygupta@gmail.com>

Constrained regression is used to avoid overfitting.
Cost Function J -
unconstrained regressions - linear regression
constrained - ridge/lasso
J(w0, w1, w2.. wn) = Cost Function over multiple variables (variables can be sq or cube etc.)
constraint may be - Sigma w-square < something (Ridge) or Sigma w < something (Lasso)
This is also called Regularisation - Introducing additional term to penalize large weights

MSE = RSS/n
R2 = 1- RSS/TSS
RMSE = Sqroot(MSE)
RSE = Sqroot(RSS/n-2)
Adjusted R2 = If more noisy variables come then R2 goes down
[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Mon, Sep 3, 2018 at 8:47 AM
To: Sumit Gupta <montygupta@gmail.com>

Hypothesis testing says that high p value means that the null hypothesis is true.. so in regression that null hypothesis is the value of coeffecient of a variable is 0. So high p value means that we can remove the variable.
[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Wed, Sep 5, 2018 at 9:10 AM
To: Sumit Gupta <montygupta@gmail.com>

Null Hypothesis (H0) - = or >= or <=
Alternate Hypothesis (H1) - != or > or <
[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Wed, Sep 5, 2018 at 10:27 AM
To: Sumit Gupta <montygupta@gmail.com>

p-value -

Suppose avg height of population is 5.7 and the samples taken has avg of 5.9 and the p-value comes to 0.05
If we say that the null hypothesis is that the height of an individual is less than 5.9 and if we take a sample again then only 5 people of 100 will have height above 5.7 and that means that we accept or we fail to reject the null hypothesis. If the p-value goes higher then the likely hood of rejecting null hypothesis is higher. The comparison is made using the significance level.

The amount of paracetamol deemed safe by the drug regulatory authorities is 500 mg
Upon sampling 900 tablets, you get an average content of 510 mg with a standard deviation of 110. What does the test suggest, if you set the significance level at 5%? Should you be happy with the manufacturing process or should you ask the production team to alter the process? Is it a regulatory alarm or a quality issue?

You can calculate the z-score for the sample mean 510 mg using the formula: $(\bar{x} - \mu) / (\sigma / \sqrt{N})$. This gives you (510 - 500) / (110 / $\sqrt{900}$) = (10) / (110 / 30) = 2.73. Notice that, since the sample mean lies on the right side of the hypothesised mean of 500 mg, the z-score comes out to be positive.

*The value in the z-table corresponding to 2.7 on the vertical axis and 0.03 on the horizontal axis is 0.9968. Since the sample mean is on the right side of the distribution and this is a two-tailed test (because we want to test if the value of the paracetamol is too low or too high), the p-value would be 2 * (1 - 0.9968) = 2 * 0.0032 = 0.0064*

Here, the p-value comes out to be 0.0064. Since the p-value is less than the significance level (0.0064 < 0.05) and smaller p-value gives you greater evidence against the null hypothesis. So you reject the null hypothesis that the average amount of paracetamol in medicines is 500 mg. So, this is a regulatory alarm for the company and the manufacturing process needs to change.

[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Fri, Sep 7, 2018 at 10:06 AM
To: Sumit Gupta <montygupta@gmail.com>

Logistic Regression

Sigmoid curve - 1/1+e raised - (Bo + Bix)
likelihood function - (1-p1)(1-p2)p3(1-p4) - if p1 is No and p3 is yes
GLM - Generalized linear model (coeff is B0 and features will be Bix)
The above eq for sigmoid curve is complex. So we use odds which is P/1-P (prob of something divided by prob of something not). So a value of 1 is equal prob of both.
plot ln(odds) = log (Bo + Bix)
linear effect on X has multiplicative effect on odds
telecom['PhoneService'] = telecom['PhoneService'].map({'Yes':1, 'No':0});

Standardisation x = x - mean(x)/ sd(x)
Normalisation x = x - mean(x)/ max(x) - min(x)

Standardisation is preferred for PCA as it compares similarities among features.
Normalisation is helped to fit a range example image processing. range of 0 to 255 RGB color range

predict_prob will give probability which can be used to get confusion matrix

confusion matrix
accuracy = all correct predicted (y and n)/ all total
sensitivity = all correct yes predicted/ all yes
specificity = all correct no predicted/ all no

true negative , false positive
false negative, true positive
(positive and negative are predictions)

[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Sat, Sep 8, 2018 at 1:23 PM
To: Sumit Gupta <montygupta@gmail.com>

Advanced Linear Regression

Linear word means that coefficients are linear and not the features

[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Sat, Sep 8, 2018 at 1:34 PM
To: Sumit Gupta <montygupta@gmail.com>

Coeffs = (X transpose * X)inverse * X transponder * all y observations

X - n*k
n all data points
K all derived features

[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Sat, Sep 8, 2018 at 1:46 PM

To: Sumit Gupta <montygupta@gmail.com>

Identity Matrix - 1 only on the diagonal

Pipelining in the Python
It is crucial in machine learning modelling that the data you feed the model is specifically pre-processed and refined for the problem you want to solve. This includes data cleaning, scaling, imputation, preprocessing, feature engineering, and so on.

Scikit-learn "Pipeline" is a utility that provides a way to automate a machine learning workflow. It works by allowing several preprocessing steps to be chained together. In the scikit-learn library, Pipelines help to clearly define and automate these standard workflows. Refer the additional references section to read more about pipelines in scikit-learn.
[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Sat, Sep 8, 2018 at 2:01 PM
To: Sumit Gupta <montygupta@gmail.com>

Ridge - Add sum of squares of coefficients as a regularisation - square ridge
Lasso is sum of mod of coefficients
[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Sat, Sep 8, 2018 at 2:18 PM
To: Sumit Gupta <montygupta@gmail.com>

Ridge -
Coeffs = (X transpose * X + lambda*Identity matrix) inverse * X transponder * all y observations

This helps with regularisation and also helps inverse the matrix

Lasso is computationally more intensive but it's very good as it removes some features by making the coefficients as 0
[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Sat, Sep 8, 2018 at 4:51 PM
To: Sumit Gupta <montygupta@gmail.com>

SVM

Basically shortest distance from the middle plane

As discussed in the lectures, the following two constraints determine the equation of the maximal margin classifier:
Sigma wi 2 = 1 - to normalise to 1
Predict (1 or -1) * Dot Product wi and ui >= M minimum distance
[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Sat, Sep 8, 2018 at 5:21 PM
To: Sumit Gupta <montygupta@gmail.com>

The Soft Margin Classifier overcomes the drawbacks of the Maximal Margin Classifier by allowing certain points to be misclassified. You control the amount of misclassifications using the cost of misclassification 'C', where C is the maximum value of the summation of the slack variable epsilon(ϵ), i.e.

If C is high, a higher number of points are allowed to be misclassified or violate the margin. In this case, the model is flexible, more generalisable, and less likely to overfit. In other words, it has a high bias.

On the other hand, if C is low, a lesser numer of points are allowed to be misclassified or violate the margin. In this case, the model is less flexible, less generalisable, and more likely to overfit. In other words, it has a high variance.

So, C represents the 'liberty of misclassification' that you provide to the model.

Note that the C defined above and the parameter C used in the SVC() function in python are the inverse of each other. In SVC(), C represents the 'penalty for misclassification'.
[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Sat, Sep 8, 2018 at 5:39 PM
To: Sumit Gupta <montygupta@gmail.com>

Linear Kernel, Polynomial Kernel and RBF kernel
Kernels covert non linear problem to linear problem
[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Sat, Sep 8, 2018 at 6:27 PM
To: Sumit Gupta <montygupta@gmail.com>

Decision Tree
Gini Index - sum of square of prob for each feature to get a gini index and the hugest is used to partition. This is homogeneity measure

Entropy and Information gain - Entropy of D - Entropy is Da. Higher it is the better it is
D = - sigma pi log pi
Da = sigma a/Total * pi log pi

Splitting by R2 - R2 of partitions should not be less than the main node. Usually helpful for regression
[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Sat, Sep 8, 2018 at 6:34 PM
To: Sumit Gupta <montygupta@gmail.com>

There are two ways to control overfitting in trees:
Truncation - Stop the tree while it is still growing so that it may not end up with leaves containing very few data points.
Pruning - Let the tree grow to any complexity. Then, cut the branches of the tree in a bottom-up fashion, starting from the leaves. It is more common to use pruning strategies to avoid overfitting in practical implementations.
[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Sun, Sep 9, 2018 at 4:07 AM
To: Sumit Gupta <montygupta@gmail.com>

Bagging

Bootstrapped Aggregation

the OOB error is calculated as the number of observations predicted wrongly as a proportion of the total number of observations.
[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Sun, Sep 9, 2018 at 4:28 AM
To: Sumit Gupta <montygupta@gmail.com>

Clustering -
So to compute silhouette metric, avg distance from own cluster - cohesion and avg distance from other clusters - separation
Hopkins Teat - Whether the data is good for clustering

RFM analysis
In RFM analysis, you look at the recency, frequency and the monetary scores of all the customers for segmentation.
Recency: It measures how recently you visited the store or made a purchase
Frequency: It measures the frequency of the transactions the customers made
Monetary: It measures how much the customer spent on purchases he/she made

However, in this plot (Fig 1), you can notice a distinct elbow. Beyond the elbow point, the additional (marginal) decrease in inertia with each increase in the cluster number is not very prominent. Thus, the elbow in the curve gives an estimate of the optimal number K in K Means.

In other words, inertia represents, how tightly the different clusters are formed. As we increase the number of clusters, the inertia value is bound to decrease as the individual clusters become more compact. Thus, the plot of inertia against the number of clusters becomes a monotonically decreasing plot.

Let's recall what you have learnt in this session so far. You learnt about another clustering technique called Hierarchical clustering. You saw how it is different from K-Means clustering. One major advantage is that you do not have to pre-define the number of clusters. However, since you compute the distance of each point from every other point, it is time-consuming and needs a lot of processing power.

[Quoted text hidden]

Sumit Gupta <montygupta@gmail.com> Sun, Sep 9, 2018 at 4:46 AM
To: Sumit Gupta <montygupta@gmail.com>

K Mode and K prototype clustering
K mode is for categorical variables
K prototype is for continuous and categorical variables
DB Scan clustering
Density based and uses Euclidean distance. Requires two params - EPS or min samples
EPS is the radius
DB scan used by Netflix
GMM - Gaussian Mixture Model. Uses soft clustering. Used in voice recognition etc

[Quoted text hidden]