Friday, August 18, 2017 9:28 PM

A test has a true positive rate of 100% and false positive rate of 5%. There is a population with a 1/1000 rate of having the condition the test identifies. Considering a positive test, what is the probability of having that condition?

Let's suppose you are being tested for a disease, if you have the illness the test will end up saying you have the illness. However, if you don't have the illness- 5% of the times the test will end up saying you have the illness and 95% of the times the test will give accurate result that you don't have the illness. Thus there is a 5% error in case you do not have the illness.

Out of 1000 people, 1 person who has the disease will get true positive result.

Out of the remaining 999 people, 5% will also get true positive result.

Close to 50 people will get a true positive result for the disease.

This means that out of 1000 people, 51 people will be tested positive for the disease even though only one person has the illness. There is only a 2% probability of you having the disease even if your reports say that you have the disease.

How can you deal with different types of seasonality in time series modelling?

Seasonality in time series occurs when time series shows a repeated pattern over time. E.g., stationary sales decreases during holiday season, air conditioner sales increases during the summers etc. are few examples of seasonality in a time series.

Seasonality makes your time series non-stationary because average value of the variables at different time periods. Differentiating a time series is generally known as the best method of removing seasonality from a time series. Seasonal differencing can be defined as a numerical difference between a particular value and a value with a periodic lag (i.e. 12, if monthly seasonality is present)

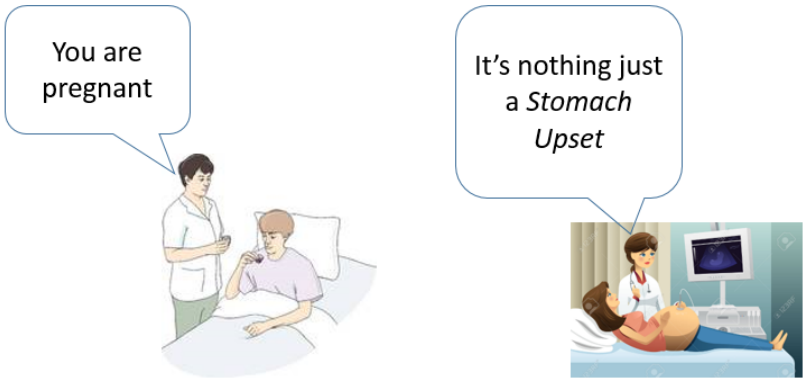
Can you cite some examples where a false positive is important than a false negative?

Before we start, let us understand what are false positives and what are false negatives.

False Positives are the cases where you wrongly classified a non-event as an event a.k.a Type I error.

And, False Negatives are the cases where you wrongly classify events as non-events, a.k.a Type II error.

False Positive False Negative



In medical field, assume you have to give chemo therapy to patients. Your lab tests patients for certain vital information and based on those results they decide to give radiation therapy to a patient.

Assume a patient comes to that hospital and he is tested positive for cancer (But he doesn't have cancer) based on lab prediction. What will happen to him? (Assuming Sensitivity is 1)

One more example might come from marketing. Let's say an ecommerce company decided to give $1000 Gift voucher to the customers whom they assume to purchase at least $5000 worth of items. They send free voucher mail directly to 100 customers without any minimum purchase condition because they assume to make at least 20% profit on sold items above 5K.

Now what if they have sent it to false positive cases?

What do you understand by statistical power of sensitivity and how do you calculate it?

Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, RF etc.). Sensitivity is nothing but "Predicted TRUE events/ Total events". True events here are the events which were true and model also predicted them as true.

Calculation of seasonality is pretty straight forward-

Seasonality = True Positives /Positives in Actual Dependent Variable

Where, True positives are Positive events which are correctly classified as Positives.

What does P-value signify about the statistical data?

P-value is used to determine the significance of results after a hypothesis test in statistics. P-value helps the readers to draw conclusions and is always between 0 and 1.

- P- Value > 0.05 denotes weak evidence against the null hypothesis which means the null hypothesis cannot be rejected.
- P-value <= 0.05 denotes strong evidence against the null hypothesis which means the null hypothesis can be rejected.
- P-value=0.05is the marginal value indicating it is possible to go either way.

What is logistic regression? Or State an example when you have used logistic regression recently.

Logistic Regression often referred as logit model is a technique to predict the binary outcome from a linear combination of predictor variables. For example, if you want to predict whether a particular political leader will win the election or not. In this case, the outcome of prediction is binary i.e. 0 or 1 (Win/Lose). The predictor variables here would be the amount of money spent for election campaigning of a particular candidate, the amount of time spent in campaigning, etc.

What are Recommender Systems?

A subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product. Recommender systems are widely used in movies, news, research articles, products, social tags, music, etc.

What is multicollinearity and how you can overcome it?

In regression, "multicollinearity" refers to predictors that are correlated with other predictors. Multicollinearity occurs when your model includes multiple factors that are correlated not just to your response variable, but also to each other. In other words, it results when you have factors that are a bit redundant.

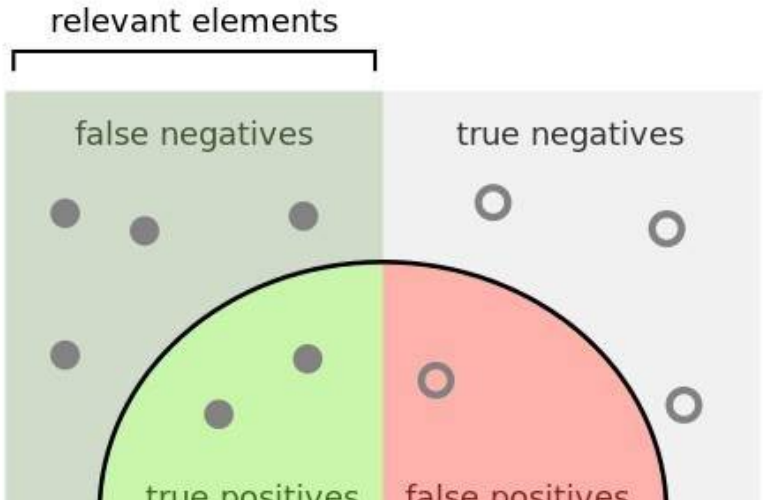
problem with multicollinearity.

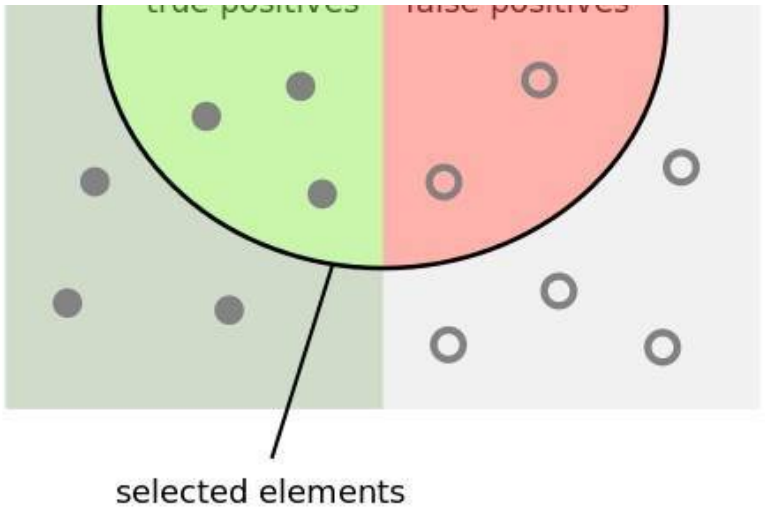
Multicollinearity increases the standard errors of the coefficients. Increased standard errors in turn means that coefficients for some independent variables may be found not to be significantly different from 0. In other words, by overinflating the standard errors, multicollinearity makes some variables statistically insignificant when they should be significant. Without multicollinearity (and thus, with lower standard errors), those coefficients might be significant.

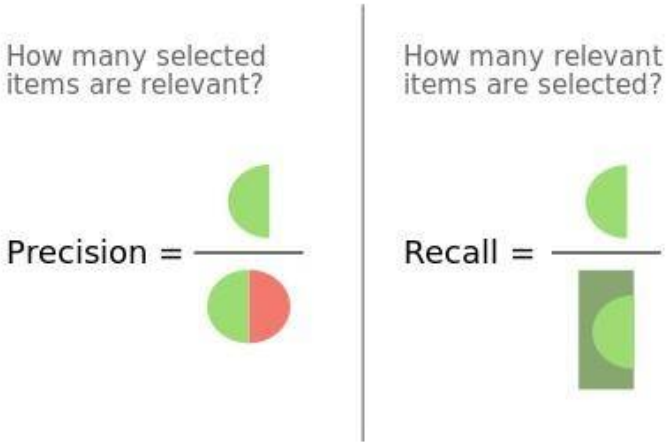
How Can I Deal With Multicollinearity?

Remove highly correlated predictors from the model
Use PCA or dimensionality reduction techniques

Recall and Precision







Explain the steps in making a decision tree.

Answer:

1. Take the entire data set as input.
2. Look for a split that maximizes the separation of the classes. A split is any test that divides the data in two sets.
3. Apply the split to the input data (divide step).
4. Re-apply steps 1 to 2 to the divided data.
5. Stop when you meet some stopping criteria.
6. This step is called pruning. Clean up the tree if you went too far doing splits.

Explain cross-validation.

Answer:

It is a model validation technique for evaluating how the outcomes of a statistical analysis will generalize to an independent data set. It is mainly used in backgrounds where the objective is forecast and one wants to estimate how accurately a model will accomplish in practice. The goal of cross-validation is to term a data set to test the model in the training phase (i.e. validation data set) in order to limit problems like overfitting, and gain insight on how the model will generalize to an independent data set.

What's the trade-off between bias and variance?
Bias is error due to erroneous or overly simplistic assumptions in the learning algorithm you're using. This can lead to the model underfitting your data

Variance is error due to too much complexity in the learning algorithm you're using, which can lead your model to overfit the data. You'll be carrying too much noise from your training data for your model to be very useful for your test data.

How is KNN different from k-means clustering?

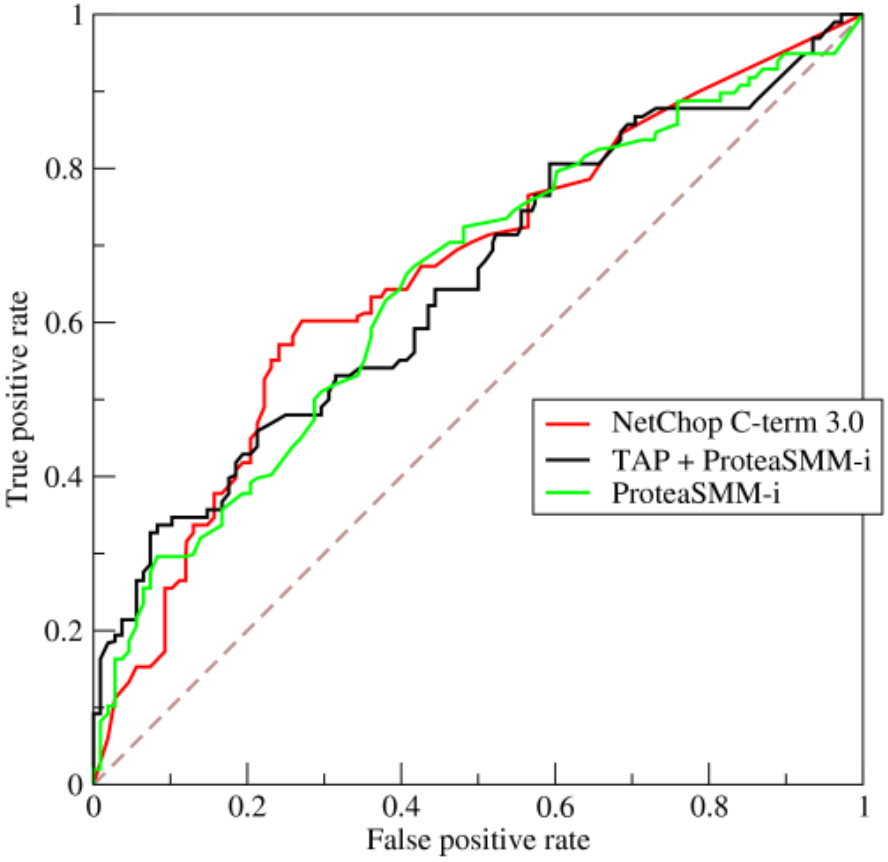
More reading: How is the k-nearest neighbor algorithm different from k-means clustering? (Quora)

K-Nearest Neighbors is a supervised classification algorithm, while k-means clustering is an unsupervised clustering algorithm.

Explain how a ROC curve works.

More reading: Receiver operating characteristic (Wikipedia)

The ROC curve is a graphical representation of the contrast between true positive rates and the false positive rate at various thresholds. It's often used as a proxy for the trade-off between the sensitivity of the model (true positives) vs the fall-out or the probability it will trigger a false alarm (false positives).



Name an example where ensemble techniques might be useful.

More reading: Ensemble learning (Wikipedia)

Ensemble techniques use a combination of learning algorithms to optimize better predictive performance. They typically reduce overfitting in models and make the model more robust (unlikely to be influenced by small changes in the training data).

You could list some examples of ensemble methods, from bagging to boosting

What's the F1 score? How would you use it?

More reading: F1 score (Wikipedia)

The F1 score is a measure of a model's performance. It is a weighted average of the precision and recall of a model, with results tending to 1 being the best, and those tending to 0 being the worst. You would use it in classification tests where true negatives don't matter much.

What is the Box-Cox transformation used for?

The Box-Cox transformation is a generalized "power transformation" that transforms data to make the distribution more normal.

For example, when its lambda parameter is 0, it's equivalent to the log-transformation.

It's used to stabilize the variance (eliminate heteroskedasticity) and normalize the distribution.

Explain Latent Dirichlet Allocation (LDA).

Latent Dirichlet Allocation (LDA) is a common method of topic modeling, or classifying documents by subject matter.

LDA is a generative model that represents documents as a mixture of topics that each have their own probability distribution of possible words.

The "Dirichlet" distribution is simply a distribution of distributions. In LDA, documents are distributions of topics that are distributions of words.