MITOCW | MIT15_071S17_Session_6.4.09_300k

In this video we will compare all the different methods we have seen so far in this course and review what they are used for, their benefits, and limitations.

Linear regression is used to predict a continuous outcome.

Linear regression is simple and commonly used, and it works on small and large data sets.

The downside is that it assumes a linear relationship.

If we have a nonlinear relationship, we need to add variables to our analysis.

For instance, suppose y = a*log(X)+b, where x is data, and y is what we need to predict.

To be able to find the coefficients a and b through linear regression, we need to view log(X) as a new variable.

Remember that we did this in the Google homework problem.

Logistic regression is used to predict a categorical outcome.

We mainly focused on binary outcomes, like yes or no, sell or buy, accept or reject, and so on.

We have seen it applied to predict the quality of care, good or bad; the winner of the US presidential election, Republican or Democrat; parole violation and loan payment, yes or no.

In addition to its relative simplicity, logistic regression computes probabilities that can be used to assess the confidence of our prediction.

The downside is again similar to that of linear regression.

In the trees week we learned CART, which is used to predict a categorical outcome, with possibly more than two categories, like quality rating, from one to five, and three decisions, say, buy, sell, or hold.

It can also predict a continuous outcome, such as salary or price.

We have seen it applied to predict life expectancy, earnings from census data, and letter recognition.

The power of CART lies in the fact that it can handle nonlinear relationships between variables.

The tree representation makes it easy to visualize and interpret the results.

The downside is that CART may not work very well on small data sets.

Random forest is also used to predict categorical outcomes or continuous outcomes.

Its benefit over CART is that it can improve the prediction accuracy.

However, we need to adjust many parameters and it's not as easy to explain as CART This week, we learned hierarchical clustering, which is used to find similar groups.

An important aspect of clustering data into smaller groups is that we can improve our prediction accuracy by applying our predictive methods, like logistic regression for instance, on each cluster.

We expand on this cluster-then-predict idea in one of our homework problems.

Hierarchical clustering is an attractive technique, because we do not need to select the number of clusters before running the algorithm.

Also, we can visualize the clusters using a dendrogram.

The drawback though, is that hierarchical clustering is hard to use on large data sets, because of the pairwise distance calculation, as we saw in this recitation.

An alternative method is k-means clustering, which works well on data sets of any size.

However, k-means requires selecting the number of clusters before running the algorithm.

This may not be a limitation if we have an intuition of the number of clusters we want to look at, as in the medical image segmentation example.

I hope that this quick review gave you a good refresher before the competition week.

Good luck.