Definitions, Nearest Neighbors

Tuesday, April 16, 2019 12:26 PM

Instructor: Tina Peters
Book: Statistics and Data Mining, Machine Learning in Astro

Definitions:

Learning Problem:

Set of N Samples of data to predict properties of unknown data

Ex: Based on the netflix movies you currently watch, what would you watch in the future?

Sample:

Set of data collected from statistical population by defined procedure.
e.g.: random, complete, flux limited, supervised, unsupervised

Supervised

Data set includes additional properties we want to predict.

Classification (discrete)
Sample is drawn from 2 or more classes that include labels

Regression (continuous)
Sample has an attribute that is a continuous variable that we want to predict.

Unsupervised

Data has features but no labels
Goal is to discover similarities in the sample, determine the distribution or simply for visualization.

Feature

A measured attribute used in learning

Training Set - Labeled Data

Data set on which we learn

Confusion Matrix

		Classifier Prediction	
		Positive	Negative
Actual Value	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Diagnostics

Completeness

$$\frac{True\ Positive}{(True\ Positive + False\ Negative)}$$

Purity

$$\frac{True\ Positive}{(True\ Positive + False\ Positive)}$$

If you do perfectly you're diagonalizing the confusion matrix.
These diagnostics tell you how off diagonal you are.

Test Set

Data set on which we learn the learned properties.

Nearest Neighbor Classification:

Short comings
- Not Generalized
- Doesn't make a model
- Computationally Expensive

Data set of N points with D features.

$$x_{\{1,\dots,n\}} = [\leftarrow D \rightarrow]$$

$$y = \begin{matrix}\uparrow \\ N \\ \downarrow\end{matrix}$$

Supervised integers, number of classes.

We want to find the closest point in x and assign it the label of that point.

Find the **Euclidean distance** between things of a certain label.

$$D(x, x_i) = \sqrt{}\left\{ \Sigma^D \left(X_D - X_{\{i,D\}}\right)^2 \right\}$$

K is a tunable parameter
It becomes smoother the higher the value it is.

As you increase k you suppress the noise and ignore outliers.

You can give weights based on things like your errorbars.

Deterministic Classification

K nearest neighbor
Majority of k nearest neighbors

Probabilistic Classification

Ratio of the k nearest neighbors

Training Data

K fold
Split up your known data into k chunks to test on.
Cross Validation