141A lecture 11/29

INPUT: Television Budget Ratio Print (Independent variables Covariates Predictors Features)

OUTPUT: Sales (Dependent variable Response label (categorical))

We usually use modeling to estimate f (the relationship)

Parameter:

Classification: logistic regression , linear discrimination analysis

Regression: linear regression

Non-parameter :

Classification: classification tree, KNN, neural

Regression: regression tree, KNN, neural

Supervised model:

In a supervised model, we have a response Y and inputs X1,X2

Unsupervised model:

In a unsupervised model, there is no response Y. All we have are inputs X1,X2.

**KNN model**

we’ll look at the classification case.

Y has classes 1,2

X1, X2 are continuous.

K is the number of the neighbors to consider when we predict for new data.

Distance

Euclidean distance in d dimensions

Manhattan distance

12/4/2018

Introduction to statistical learning

KNN

We can adjust k , the distance metric.

As k increases, KNN tends to choose the majority class for all predictions.

As k increases, the model becomes more biased.

Bias - variance trade off

When k = 1

Small changes to the new point lead to big changes in our prediction , k = 1 is high variance.

How to implement KNN

12/6/2018

CV (M - FOLD)

Goal: estimate model errors we tend fitting with them and then predicting for the train data. This understands the error train on 1, predict on 2.

We get an error estimate zi, train on 2, predict on 1, we get an error estimate ei hat.

How to implement CV?

INPUTS: training data, m, model

OUTPUTS: error estimate

Steps for CV:

1. We shuffle the observations
2. We split the observations to m groups of approximately the same size

Sample() seq() rep() split()

1. For each fold, predict and compute an error estimate, using the other folds as training data set. Lapply()
2. Compute the average of the estimates. Mean()