Data Mining, CSCI 347, Fall 2019
Linear Models, Oct. 4

Linear models

- Use when working with numeric attributes.

Linear regression versus logistic regression

- Linear regression – used when the dependent variable is continuous and nature of the regression line is linear.
- Logistic regression - used when the dependent variable is binary in nature.

Regression – returning to a previous state
Regression for data mining – predicting a numeric quantity

Linear regression – the output is a linear combination of the weighted attributes and a bias

Given n attributes $a_1, \ldots, a_n$, the numeric class value is estimated:

$$x = w_0 + w_1 * a_1 + w_2 * a_2 + \ldots + w_n * a_n$$

where the weights $w_0, \ldots, w_n$ are calculated from the training data. The weight w_0 is called the "bias".

When talking about an instance k, can write

$$x^{(k)} = w_0 + w_1 * a_1^{(k)} + w_2 * a_2^{(k)} + \ldots. + w_n * a_n^{(k)}$$

for the class value $x^{(k)}$ given the attribute values $a_1^{(k)}, \ldots a_n^{(k)}$

Often write this as:

$$x^{(k)} = w_0 * a_0^{(k)} + w_1 * a_1^{(k)} + w_2 * a_2^{(k)} + \ldots. + w_n * a_n^{(k)}$$

where a_0 is assumed to be 1 so can concisely expressing using the summation notation, Σ, as

$$x^{(k)} = \sum_{i=0}^{n} w_i * a_i^{(k)}$$

Goal: Choose the weights $w_0, \ldots, w_n$ to minimize the sum of the squares of the differences between the actual and predicted values. That is, minimize:

$$\sum_{j=1}^{m} \left(x^{(j)} - \sum_{i=0}^{n} w_i * a_i^{(j)} \right)^2$$

where m is the number of instances in the dataset, n is the number of attributes, $x^{(j)}$ is the actual value of the jth instance, w_i is the weight of the ith attribute (except w_0 is the bias) and $a_i^{(j)}$ is the value of the ith attribute in the jth instance.

Minimizing the squares of the difference exaggerates large differences. Instead could minimize the absolute value of the differences.

Linear regression works well if the data truly is linear or near linear. Statistics offers many other regression techniques for when the data isn't linear.

Trees for Numeric Prediction

- Regression tree
  - "Decision tree" where each leaf predicts a numeric quantity
  - Leaf is the average of the class value for all instances that reach the leaf
- Model tree
  - "Regression tree" with linear regression models (formulas) at the leaf nodes

Linear Models for Classification

Multiresponse linear regression
Do for each class value

1. Take the data set and make class value 1 in all of the instances of the class, and 0 for the other records
2. Apply linear regression

This gives n formulas. Given a new instance to classify, substitute values into all n formulas and select the class when the result is the largest.

2 drawbacks to multiresponse linear regression:

1. Membership values aren't probabilities since the can fall outside the [0, 1] range.
2. Least squares regression assumes errors are independent and normally distributed with the same standard deviation. This is blatantly violated with multiresponse linear regression because the observations are only 0 and 1.