Class Notes

Statistical Computing & Machine Learning

Class 7

Professional contacts

The upcoming AI, Machine learning, and robotics meet up. Wednesday 24 Feb.

Coefficients as quantities

Coefficients in linear models are not just numbers, they are physical quantities with dimensions and units.

- Dimensions are always (dim of response)/(dim of this term)
- The model doesn't depend on the units of these quantities. The units only set the magnitude to the numerical part of the coefficient, but as a quantity a coefficient is the same thing regardless of units.
- Conversion from one unit to another by multiplying by 1, but expressed in different units, e.g. 60 seconds per minute, 2.2 pounds per kilogram.

K-nearest neighbors

K-nearest neighbors is a simple, general kind of function-building method. But some problems:

- Interpretability: but you can always take partial derivatives.
- When you have prediction (aka "explanatory") variables in dollars and in miles, how do you calculate the distance between points? What are the dimensions of distance?
  - Dimensionality refers to the physical feature, e.g. time, distance, area, volume, money, charge, luminance, mass, ...
  - Units are the ways in which dimensions are measured, e.g., cups, gallons, liters ... all refer to volume
    - Give some examples of units for each of the dimensions.
    - Some everyday quantities are dimensionless, e.g. pure numbers. Give some examples: ... (angles, percent, fractions, ... but not ratios in general.)
  - Regression fixes units automatically, since the coefficients themselves have dimensionality. They will adjust automatically to changes in units, so the model is the same regardless of whether we use miles, km, parsecs, ...

  - In KNN, to avoid dependence on units, need to do some standardization by dividing by something in the same units, e.g. sd.

- Curse of dimensionality. Let's create 1000 randomly placed points in the unit square:

```
rpts <- matrix(runif(2 * 1000), ncol = 2)
```

What's the distribution of distances from a single random point to the 1000 others:

```
our_point <- runif(2)
```

The distance between our point and each of the others

```
tmp <- matrix(our_point, ncol = 2, nrow = 1000,
    byrow = TRUE)
delta <- sqrt(rowSums((rpts - tmp)^2))
```

- How far away is a typical point?
- Write a function that takes the matrix of points and the "our point" and finds the distance from our point to each and every one of the points in the matrix.
- How far away is a typical point in 1-dimensional space?
- In 10-dimensional space?
- In 100-dimensional space?

Bayes and likelihood

Conditional probability

What we want is $p(\text{state of world}|\text{observations})$. I'll write this $p(\theta|\mathcal{O})$

Tree with cancer (benign or malignant) and cell shape (round, elongated, ruffled)

SPACE FOR THE TREE

SEE PAPER NOTES. (remember to transcribe them here)

, e.g. observe ruffled, what is the chance that the tumor is malignant.

Of the 10000 people in the study,

* 7000 had benign tumors of which 10% or 700 had ruffled cells * 3000 had malignant tumors of whom 60% or 1800 had ruffled cells

So, of the 2500 people with ruffled cells, 1800 had malignant tumors. $p(\theta|\mathcal{O})$

In-class programming activity

Introduce exponential probability distribution. Example: When will the next earthquake happen when the historical record shows a 100-year average time between earthquakes? link to day 7 activity