6.S085 Statistics for Research Projects **IAP 2014**

Lecture 5: January 24

Lecturer: Ramesh Sridharan and George Chen *Notes by: William Li*

5.1 Logistic Regression

Last time:

Model: $y = X\beta + \epsilon$

This is good if there is a linear relation between y and columns of X

What if y has a non-linear pattern?

Generalized Linear Models

$y = g^{-1}(X\beta) + \epsilon$ (noise doesn't have ot be independent anymore

g^{-1} — nonlinear "link function"

One of the most common forms: logistic regression

$g^{-1}(z) = \frac{1}{1+\exp(-z)}$

Logistic regression is useful when output data (y) is binary, or between 0 and 1

5.2 Non-Parametric/Distribution-Free Statistics

Last week: t-test and variances; we assumed normality

In general, assumed known distribution, computed p-values/confidence intervals

If we had a sample mean with unknown variances — t distribution

Review of p-value: probability of observed statistic or something more extreme if null hypothesis is true (t value is usually the threshold, p value is the probability mass)

5.2.1 Comparing distributions

Kolmogorov-Smirnov Test

- Compare two distributions
- Do two distributions look almost the same, or are they different?

- Based on cumulative distribution functions (CDFs) For random value x, $F(a) = P(x < a)$
- Empirical CDF: For any value a, what % of data is $\leq a$? (the empirical CDF is a property of the data)

Find the biggest difference in the CDF

$D = \max_x |F_1(x) - F_2(x)|$

Theoretical result: if $F_1$is empirical CDF of data generated from F_2, $\lim_{n\to\infty} D = 0$

The Kolmogorov-Smirnov statistic is sensitive to any difference (your software package will compute it for you)

If you want to compare your data vs. normal distribution, use Shapiro-Wilk test (uses quantiles of your data and quantiles of the normal distribution)

5.2.2 Wilcoxon Signed-Rank Test

For comparing medians of two distributions (medians are less sensitive to outliers, e.g. 1,2,4,4,8,200)

For matched pairs — two datasets where there is a correspondence, often a before/after (weight, test scores, etc.)

For each pair, d_1, $S_i \in \{(\pm 1)\}$

Rank all d_i, compute R_i (smallest to largest)

Example ranking (smallest to largest): d_5, d_3, d_1, d_4, d_2

$R_5 = 1, R_3 = 2, ...$

$W = |\sum_i R_i|$

W has a known distribution (you can compute p-values, confidence intervals, etc. on this)

W is a measure of how different the medians are; if they are almost the same, the terms will cancel out and be zero

Example calculation will look like $-1 * 1 + 1 * 2...$

If W is large, medians are different

Mann-Whitney U Test: Similar, but doesn't require matched pairs

5.3 Resampling Methods

What if the test statistical distribution is unknown?

Key idea: use the data to tell us about the distribution

5.3.1 Permutation Tests

Used for hypothesis testing

Used when we don't have a null hypothesis

Comparing statistic across two groups

Idea: is there anything special about the way we labeled the groups?

Example: Consider two groups, A and B

$\bar{x}_A - \bar{x}_B = w$

How do we know whether w is big or small?

- Relabel points:
  - put all $n + m$ points together
  - pick n points, call it A_1
  - call the rest (m) points, B_1
- Compute test statistic $(w_1 = \bar{x}_{A_1} - \bar{x}_{B_1})$
- Repeat for different relabelings: $W = \{w_1, w_2, ..., w_k\}$

How unlikely is w given your set of w values, W? (Generate an empirical distribution of W)

Exact test: compare versus all relabelings

Monte Carlo approximation: randomly sampl erelabelings

5.3.2 Permutation Test: Other Examples

Time series analysis: (3,10,117,20,9)

Consider a random reordering: (11,9,10,20,7,3)

Compute the statistic for the actual and reordered points

The question we're asking: Is the ordering meaningful?

5.3.3 Detecting Dishonest Teachers

Two statistics developed:

- A: Comparing student scores to year before and year after
- B: Pattern of ABCD responses; how similar are they?

Permutation test: permute teacher and student matches and compute null distribution

For A, compute 50th-75th percentile of values

5.3.4 Bootstrap

Data are samples from the true distribution

Wouldn't it be nice if we could get another sample?

Idea: resample from data with replacement

Key insight from 1979 bootstrapping paper: random samples out of random samples → more random samples

If data has N points, resample n ($n \leq N$) points with replacement

Repeat many times; this will give you multiple datasets

Purpose: understand variability → compute confidence intervals

Compute statistic w on each dataset, $\{w_1, ..., w_n\}$

Can compute the confidence interval: perhaps find 25th to 75th percentile

In machine learning, it could be used to compute the variability of estimates

5.3.5 Jacknife

Like bootstrapping, but:

Instead of bootstrapping, use all but one point, repeat for each point

Kind of like bootstrap, with $n = N - 1$

It's exact instead of randomized

5.4 Model Selection

Last time: use Lasso to get sparsity in linear regression

Recall LASSO: $\min \sum_i (y_i - X_i\beta)^2 + \lambda \sum_k |\beta_k|$