Avoid False Discoveries

Huiping Cao

Outline

- Statistical Background
- Significance Testing
- Hypothesis Testing

Motivation (1)

- An algorithm applied to a set of data will usually produce some result(s)
  - There have been claims that the results reported in more than 50% of published papers are false. (Ioannidis)
- Results may be a result of random variation
  - Any particular data set is a finite sample from a larger population
  - Often significant variation among instances in a data set or heterogeneity in the population
  - Unusual events or coincidences do happen, especially when looking at lots of events
  - For this and other reasons, results may not replicate, i.e., generalize to other samples of data

Motivation (2)

- Results may not have domain significance
  - Finding a difference that makes no difference
- Data scientists need to help ensure that results of data analysis are not false discoveries, i.e., not meaningful or reproducible

Statistical Testing

- Statistical approaches are used to help avoid many of these problems
- Statistics has well-developed procedures for evaluating the results of data analysis
    - Significance testing
    - Hypothesis testing

Probability and Distributions

- Variables are characterized by a set of possible values
  - Called the domain of the variable
  - Examples
    - True or False for binary variables
    - Subset of integers for variables that are counts, such as number of students in a class
    - Range of real numbers for variables such as weight or height
- A probability distribution function (PDF) describes the relative frequency with which the values are observed
- Call a variable with a distribution a random variable

Probability and Distributions (cont.)

- For a discrete variable we define a probability distribution by the relative frequency with which each value occurs
  - Let X be a variable that records the outcome flipping a fair coin: heads (1) or tails (0)
  - $P(X=1) = P(X=0) = 0.5$ (P stands for "probability")
  - If f is the distribution of X, $f(1) = f(0) = 0.5$
- Probability distribution function has the following properties
  - Minimum value 0, maximum value 1
  - Sums to 1, i.e., $\sum_{all\ values\ of\ X} f(x) = 1$

Binomial Distribution

- Number of heads in a sequence of n coin flips
  - Let R be the number of heads
  - R has a binomial distribution
  - $$P(R=k) = \binom{n}{k} P(x=1)^k P(x=0)^{n-k}$$
  - What is $P(R=k)$ given $n=10$ and $P(X=1)=0.5$?

k	*P(R = k)*
0	0.001
1	0.01
2	0.044
3	0.117
4	0.205
5	0.246
6	0.205
7	0.117
8	0.044
9	0.01
10	0.001

Probability and Distributions $\cdots$

- Probability of any specific value is 0
- Only intervals of values have non-zero probability
  - Examples: $P(X > 3)$, $P(X < -3)$, $P(-1 < X < 1)$
  - If f is the distribution of X, $P(X > 3) = \int_3^\infty f(X)dx$
- Probability density has the following properties
  - Minimum value 0
  - Integrates to 1, i.e., $\int_{-\infty}^{+\infty} f(X) = 1$

Gaussian Distribution

- The Gaussian (normal) distribution is the most commonly used

$$f(X) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Where μ and σ are the mean and standard distribution of the distribution

$$\mu = \int_{-\infty}^{+\infty} Xf(X)dx$$

$$\sigma = \int_{-\infty}^{+\infty} (X-\mu)^2 f(X)dx$$

Statistical Testing · · ·

- Make inferences (decisions) about the validity of a result
- For statistical inference (testing), we need two things:
  - A statement that we want to disprove
    - Called the **null hypothesis** (H_0)
    - The null hypothesis is typically a statement that the result is merely due to random variation
    - It is typically the opposite of what we would like to show
  - A random variable, R, called a **test statistic**, for which we know or can determine a distribution if H_0 is true.
    - The distribution of R under H_0 is called the null distribution
    - The value of R is obtained from the result and is typically numeric

Examples of Null Hypotheses

- A coin or a die is a fair coin.
- The difference between the means of two samples is 0.
- The purchase of a particular item in a store is unrelated to the purchase of a second item, e.g., the purchase of bread and milk are unconnected.
- The accuracy of a classifier is no better than random.

Significance Testing

- Significance testing was devised by the statistician Fisher.
- Only interested in whether null hypothesis is true.
- For many years, significance testing has been a key approach for justifying the validity of scientific results.
- Introduced the concept of p-value, which is widely used.

How Significance Testing Works

- Analyze the data to obtain a **result**
  - For example, data could be from flipping a coin 10 times to test its fairness
- The result is expressed as a value of the **test statistic**, R
  - For example, let R be the number of heads in 10 flips
- Compute the **probability of seeing the current value** of R or something more extreme
  - This probability is known as the p-**value** of the test statistic

How Significance Testing Works (cont.)

- If the p-value is sufficiently small, we reject the null hypothesis, H_0 and say that the result is statistically significant
  - We say we reject the null hypothesis, H_0
  - A threshold on the p-value is called the significance level, α
  - Often the significance level is 0.01 or 0.05
- If the p-value is not sufficiently small, we say that we fail to reject the null hypothesis
  - Sometimes we say that we accept the null hypothesis but a high p-value does not necessarily imply the null hypothesis is true

Example: Testing a coin for fairness

- H_0: $P(X = 1) = P(X = 0) = 0.5$
- Define the test statistic R to be the number of heads in 10 flips
- Set the significance level α to be 0.05
- The number of heads R has a binomial distribution
- For which values of R would you reject H0?

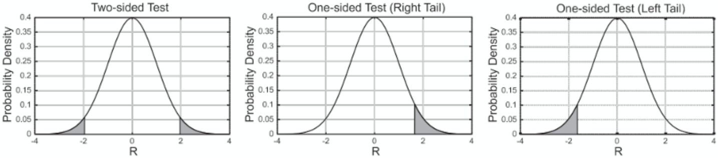
k	$P(R = k)$
0	0.001
1	0.01
2	0.044
3	0.117
4	0.205
5	0.246
6	0.205
7	0.117
8	0.044
9	0.01
10	0.001

One-sided and Two-sided Tests

- More extreme can be interpreted in different ways
- For example, an observed value of the test statistic, R_{obs}, can be considered extreme if
  - it is greater than or equal to a certain value, R_H
  - smaller than or equal to a certain value, R_L, or
  - outside a specified interval, $[R_L, R_H]$
- The first two cases are "one-sided tests" (right-tailed and left-tailed, respectively).
- The last case results in a "two-sided test".

One-sided and Two-sided Tests (cont.)

- Example of one-tailed and two-tailed tests for a test statistic R that is normally distributed for a roughly 5% significance level.



Estimating Null Distribution

- Requirement: know how the test statistic is distributed under the null hypothesis
- In conventional problems of statistical testing: this requirement is kept in mind when collecting the data.
- For example: Test effect of a new drug in curing a disease
  - Experimental data is usually collected from two groups of subjects, one group is administered the drug, other group (control group) is not.
  - The data samples from the two groups provide information to estimate the alternative and null distributions
- For many data mining problems, the observational data are collected without prior hypothesis in mind

Estimating Null Distribution for Data Mining Problems

- Generating synthetic data sets
  - For analyses involved unlabeled data:
    - Clustering
    - Frequent pattern mining
- Randomizing class labels
  - For classification
- Resampling Instances: have multiple samples from the underlying population of data instance.
  - Bootstrap sampling
  - K-fold cross validation

Randomizing Class Labels

- To generate new data:
  - Randomly permute the class labels (permutation testing)
  - The new data set is identical to the old data except for the label assignments
- A classifier is built on each of these data sets and a test statistic (e.g., accuracy) is calculated
- The resulting set of values can be used to estimate the null distribution of the test statistic

Estimate the Null Distribution of the Test Statistic

- Given multiple samples of data sets generated under the null hypothesis, compute the test statistic on every set of samples.
- Fit statistical models (e.g. the normal or the binomial distribution) on the test statistic values
- Other way is using non-parametric approaches (e.g., counting), given enough samples.

Evaluating Classification Performance

- A classifier has a testing accuracy of x%
- Validity: how likely it is to obtain x% accuracy by random chance, i.e., when there is no relationship between the attributes in the data set and the class label.
- Setup: Learn a classifier on a training set and test set
- Use a measure of the classifier's performance on the test set (e.g., precision, recall, accuracy) as the test statistic
- The null hypothesis: the classifier is not able to learn a generalizable relationship between the attributes and the class labels.

Evaluating Classification Performance - Randomization

- Generate new sample data sets under the null hypothesis that there are random relationships between the attributes and class labels by randomizing class labels on the training data
- Learn a classifier on every new sample training set
- Apply the learned models on the test set to obtain a null distribution of the test statistic
- Example: we use *accuracy* as test statistic. The accuracy for the model learned using original labels should be significantly higher than most of all of the accuracies generated by models learned over randomly permuted labels.

Statistical Testing

- One major limitation of statistical testing: it does not explicitly specify and alternative hypothesis, which is typically the statement we would like to establish as true.
  - Can be used to reject null hypothesis
  - Not suitable for determining whether an observed result actually supports alternative hypothesis.

Neyman-Pearson Hypothesis Testing

- Devised by statisticians Neyman and Pearson in response to perceived shortcomings in significance testing
  - Explicitly specifies an alternative hypothesis, H_1
  - Significance testing cannot quantify how an observed results supports H_1
  - Define an alternative distribution which is the distribution of the test statistic if H_1 is true
  - We define a critical region for the test statistic R
    - If the value of R falls in the critical region, we reject H_0
    - We may or may not accept H_1 if H_0 is rejected
  - The significance level, α, is the probability of the critical region under H_0

Hypothesis Testing (cont.)

- Type I Error (α): Error of incorrectly rejecting the null hypothesis for a result.
  - It is equal to the probability of the critical region under H_0, i.e., is the same as the significance level, α.
  - Formally, $\alpha = P(R \in \mathit{Critical\ Region}|H_0)$
- Type II Error β: Error of falsely calling a result as not significant when the alternative hypothesis is true.
  - It is equal to the probability of observing test statistic values outside the critical region under H_1
  - Formally, $\beta = P(R \notin \mathit{Critical\ Region}|H_1)$.

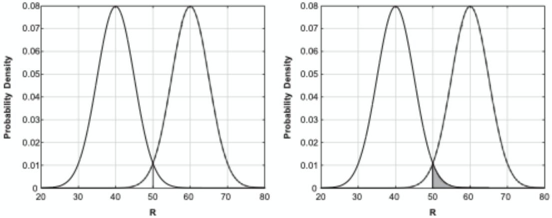
Hypothesis Testing (cont.)

- Power: which is the probability of the critical region under H_1, i.e., $1 - \beta$
  - Power indicates how effective a test will be at correctly rejecting the null hypothesis.
  - Low power means that many results that actually show the desired pattern or phenomenon will not be considered significant and thus will be missed.
  - Thus, if the power of a test is low, then it may not be appropriate to ignore results that fall outside the critical region.

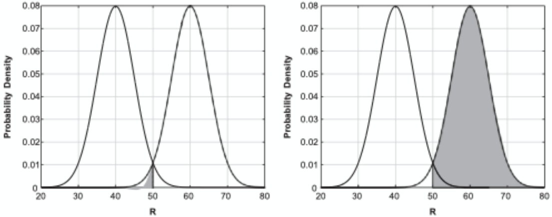
Example: Classifying Medical Results

- The value of a blood test is used as the test statistic R to identify whether a patient has a particular disease or not.
  - H_0: For patients not having the disease R has distribution $N(40, 5)$
  - H_1: For patients having the disease, R has distribution $N(60, 5)$
  - $\alpha = \int_{50}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(R-\mu)^2}{2\sigma^2}} dR = \int_{50}^{\infty} \frac{1}{\sqrt{50\pi}} e^{-\frac{(R-40)^2}{50}} dR = 0.023$, $\mu = 40$, $\sigma = 5$
  - $\beta = \int_{-\infty}^{50} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(R-\mu)^2}{2\sigma^2}} dR = \int_{-\infty}^{50} \frac{1}{\sqrt{50\pi}} e^{-\frac{(R-60)^2}{50}} dR = 0.023$, $\mu = 60$, $\sigma = 5$
  - Power $=$ 1-β $= 0.977$
  - See figures on the next page

α, β, and Power for Medical Testing Example



Distribution of test statistic for the alternative hypothesis (rightmost density curve) and null hypothesis (leftmost density curve). Shaded region in right subfigure is α.



Shaded region in left subfigure is β and shaded region in right subfigure is power.

References

- Chapter 10, Introduction to Data Mining (2nd Edition) by Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar