0.1 Statistical Data Mining

Statistical data mining, also known as knowledge or data discovery, is a computerized method of collecting and analyzing information. The data-mining tool takes data and categorizes the information to discover patterns or correlations that can be used in important applications, such as medicine, computer programming, business promotion, and robotic design.

Statistical data mining techniques use complex mathematics and complicated statistical processes to create an analysis.

Data mining involves five major steps.

1. The first data mining application collects statistical data and places the information in a warehouse-type program.
2. Next, the data in the warehouse is organized and creates a management system.
3. The next step creates a way to access the managed data.
4. Then, the fourth step develops software to analyze the data, also known as data mining regression,.
5. The final step facilitates using or interpreting the statistical data in a practical way.

Generally, data mining techniques integrate analytical and transaction data systems. Analytical software sorts through both types of data systems using open-ended user questions. Open-ended questions allow countless answers so programmers are not influencing the results of the sorting. Programmers create lists of questions to assist in categorizing the information using an overall focus.

Sorting is then based on developing classes and clusters of data, associations found in the data, and attempts to define patterns and trends based on the associations. For example, Google collects information on users' purchasing habits to assist in placing online advertising. Open-ended questions used to sort this buyer data focus on buying preferences or viewing habits of Internet users.

Computer scientists and programmers focus on the analysis of the statistical data that is collected. Creation of decision trees, artificial neural networks, nearest neighbor method, rule induction, data visualization, and genetic algorithms all use the statistically-mined data.

These classification systems assist in interpreting the associations discovered by the analytical data programs. Statistical data mining involves small projects that can be done on a small scale on a home computer, but most data mining association sets are so large and the data mining regression so complicated that they require a supercomputer or a network of high-speed computers.

Statistical data mining collects three general types of data, including operational data, non-operational data, and meta data. In a clothing store, operational data is basic data used to run the business, such as accounting, sales, and inventory control.

Non-operational data, which is indirectly related to the business, includes estimates of future sales and general information about the national clothing market.

Meta data concerns the data itself. A program using meta data might sort store customers into classifications based on gender or geographic location of the clothing buyers or the customers favorite color, if that data was collected.

0.1.1 Applications of Data Mining

A data mining application can be extremely sophisticated and the statistical data mining tool may have widespread practical applications. The study of disease outbreaks is one example. A 2000 data mining project analyzed the disease outbreak of cryptosporidium in Ontario, Canada to determine the causes of the increase in disease cases. The results of the data mining assisted in linking the bacteria outbreak to local water conditions and the lack of proper municipal water treatment. A field called "biosurveillance" uses epidemiological data mining to identify outbreaks of a single disease.

Computer programmers and designers also employ the study of probability and statistical data analysis to develop machines and computer programs. The Google Internet search engine was designed using statistical data mining. Google continues to collect and use data mining to create program updates and applications.

Supervised Learning

- **Supervised learning** is tasked with learning a function from labeled training data in order to predict the value of any valid input.
- Common examples of supervised learning include classifying e-mail messages as spam, labeling Web pages according to their genre, and recognizing handwriting.
- Many algorithms are used to create supervised learners, the most common being neural networks, Support Vector Machines (SVMs), and Naive Bayes classifiers.

Supervised and Unsupervised Learning

Supervised learning is tasked with learning a function from labeled training data in order to predict the value of any valid input.

Unsupervised Learning

- **Unsupervised learning** is tasked with making sense of data without any examples of what is correct or incorrect. It is most commonly used for clustering similar input into logical groups.
- Unsupervised learning can be used to reduce the number of dimensions in a data set in order to focus on only the most useful attributes, or to detect trends.
- Common approaches to unsupervised learning include k-Means, hierarchical clustering, and self-organizing maps.

0.2 Performance of Classification Procedure

These classifications are used to calculate accuracy, precision (also called positive predictive value), recall (also called sensitivity), specificity and negative predictive value:

- **Accuracy** is the fraction of observations with correct predicted classification

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

- **Precision** is the proportion of predicted positives that are correct

$$\text{Precision} = \text{Positive Predictive Value} = \frac{TP}{TP + FP}$$

- **Negative Predictive Value** is the fraction of predicted negatives that are correct

$$\text{Negative Predictive Value} = \frac{TN}{TN + FN}$$

- **Recall** is the fraction of observations that are actually 1 with a correct predicted classification

$$\text{Recall} = \text{Sensitivity} = \frac{TP}{TP + FN}$$

- **Specificity** is the fraction of observations that are actually 0 with a correct predicted classification

$$\text{Specificity} = \frac{TN}{TN + FP}$$

0.3 Machine learning: the problem setting

In general, a learning problem considers a set of n samples of data and try to predict properties of unknown data. If each sample is more than a single number, and for instance a multi-dimensional entry (aka multivariate data), is it said to have several variables, also known as attributes or ***features***.
We can separate learning problems in a few large categories:

- **Supervised learning**, in which the data comes with additional attributes that we want to predict. This problem can be either:

  Classification: samples belong to two or more classes and we want to learn from already labeled data how to predict the class of unlabeled data.
  An example of classification problem would be the digit recognition example, in which the aim is to assign each input vector to one of a finite number of discrete categories.

  Regression: if the desired output consists of one or more continuous variables, then the task is called regression.
  An example of a regression problem would be the prediction of the weight of a pony as a function of its age and height.

- **Unsupervised learning**, in which the training data consists of a set of input vectors x without any corresponding target values.
  The goal in such problems may be

  - to discover groups of similar examples within the data, where it is called ***clustering***,
  - to determine the distribution of data within the input space, known as ***density estimation***,
  - to project the data from a high-dimensional space down to two or thee dimensions for the purpose of visualization

0.4 Training and validation

Using Validation and Test Data

In some cases you might want to use only training and test data. For example, you might decide to use an information criterion to decide what effects to include and when to terminate the selection process. In this case no validation data are required, but test data can still be useful in assessing the predictive performance of the selected model. In other cases you might decide to use validation data during the selection process but forgo assessing the selected model on test data.

0.5 Supervised learning

Supervised learning is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a classifier (if the output is discrete, see classification) or a regression function (if the output is continuous, see regression). The inferred function should predict the correct output value for any valid input object. This requires the learning algorithm to generalize from the training data to unseen situations

in a "reasonable" way (see inductive bias).

0.6 Week 6 General Theory Topics

0.6.1 Steps in Building a Predictive Model

1. Find the right data
2. Define your error rate
3. Split data into:
    - **Training Set**
    - **Testing Set**
    - **Validation Set** (optional)
4. On the training set select predictor variables (features)
5. On the training set generate your predictive model
6. On the training set cross-validate

0.6.2 Descriptive vs Predictive Models

- A **descriptive model** is only concerned with modeling the structure in the observed data. It makes sense to train and evaluate it on the same dataset.
- The **predictive model** is attempting a much more difficult problem, approximating the true discrimination function from a sample of data. We want to use algorithms that do not pick out and model all of the noise in our sample. We do want to chose algorithms that generalize beyond the observed data. It makes sense that we could only evaluate the ability of the model to generalize from a data sample on data that it had not see before during training.
- **IMPORTANT** The best descriptive model is accurate on the observed data. The best predictive model is accurate on unobserved data.

0.6.3 Cross-Validation and Testing

- In order to build the best possible mode, we will split our training data into two parts: a training set and a test set.
- The general idea is as follows. The model parameters (the regression coefficients) are learned using the training set as above.
- The error is evaluated on the test set, and the meta-parameters are adjusted so that this cross-validation error is minimized.

0.6.4 Cross Validation

- The cross validation is often termed a jack-knife classification, in that it successively classifies **all cases but one** to develop a predictive model and then categorizes the case that was left out. This process is repeated with each case left out in turn.This is known as leave-1-out cross validation.
- This cross validation produces a more reliable function. The argument behind it is that one should not use the case you are trying to predict as part of the categorization process.

0.6.5 Error Rates

- We can evaluate error rates by means of a training sample (to construct build a model) and a test sample.
- An optimistic error rate is obtained by reclassifying the training data. (In the ***training data*** sets, how many cases were misclassified). This is known as the **apparent error rate**.
- The apparent error rate is obtained by using in the training set to estimate the error rates. It can be severely optimistically biased, particularly for complex classifiers, and in the presence of over-fitted models.
- If an independent test sample is used for classifying, we arrive at the **true error rate**.The true error rate (or conditional error rate) of a classifier is the expected probability of misclassifying a randomly selected pattern. It is the error rate of an infinitely large test set drawn from the same distribution as the training data.

0.6.6 Cross Validation

- In a prediction problem, a model is usually given a dataset of known data on which training is run (*training dataset*), and a dataset of unknown data (or *first seen data/ testing dataset*) against which testing the model is performed.
- Cross-validation is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice, with unseen data.
- The goal of cross validation is to define a dataset to "test" the model in the training phase, in order to limit problems like overfitting, give an insight on how the model will generalize to an independent data set (i.e., an unknown dataset, for instance from a real problem), etc.
- Cross-validation is important in guarding against testing hypotheses suggested by the data (called "***Type III errors***"), especially where further samples are hazardous, costly or impossible to collect

K-fold Cross Validation

- In k-fold cross-validation, the original data set is randomly partitioned into k equally sized subsamples (e.g. 10 samples).
- Of the k subsamples, a single subsample is retained as the testing data for testing the model, and the remaining k - 1 subsamples are used as training data.

- The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the test data.
- The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation.
- The advantage of this method over repeated random sub-sampling is that all observations are used for both training and testing, and each observation is used for testing exactly once.

Leave-One-Out Cross-Validation

- As the name suggests, **leave-one-out cross-validation (LOOCV)** involves using a single observation from the original sample as the validation data, and the remaining observations as the training data.
- This is repeated such that each observation in the sample is used once as the validation data.
- This is the same as a K-fold cross-validation with K being equal to the number of observations in the original sampling, i.e. **K=n**.

0.6.7 Binary Classification

Defining True/False Positives In general, Positive = identified and negative = rejected. Therefore:

- True positive = correctly identified
- False positive = incorrectly identified
- True negative = correctly rejected
- False negative = incorrectly rejected

Medical Testing Example:

- True positive = Sick people correctly diagnosed as sick
- False positive= Healthy people incorrectly identified as sick
- True negative = Healthy people correctly identified as healthy
- False negative = Sick people incorrectly identified as healthy.

0.6.8 Definitions (From Week 1)

Confusion Matrix
The confusion table is a table in which the rows are the observed categories of the dependent and the columns are the predicted categories. When prediction is perfect all cases will lie on the diagonal. The percentage of cases on the diagonal is the percentage of correct classifications.

Accuracy Rate
The accuracy rate calculates the proportion ofobservations being allocated to the **correct** group by the predictive model. It is calculated as follows:

$$\frac{\text{Number of Correct Classifications}}{\text{Total Number of Classifications}}$$

$$= \frac{TP + TN}{TP + FP + TN + FN}$$

Misclassification Rate
The misclassification rate calculates the proportion ofobservations being allocated to the **incorrect** group by the predictive model. It is calculated as follows:

$$\frac{\text{Number of Incorrect Classifications}}{\text{Total Number of Classifications}}$$

$$= \frac{FP + FN}{TP + FP + TN + FN}$$

0.6.9 Misclassification Cost

- As in all statistical procedures it is helpful to use diagnostic procedures to assess the efficacy of the analysis. We use **cross-validation** to assess the classification probability. Typically you are going to have some prior rule as to what is an **acceptable misclassification rate**.
- Those rules might involve things like, *"what is the cost of misclassification?"* Consider a medical study where you might be able to diagnose cancer.
- There are really two alternative costs.
  - The cost of misclassifying someone as having cancer when they don't. This could cause a certain amount of emotional grief. Additionally there would be the substantial cost of unnecessary treatment.
  - There is also the alternative cost of misclassifying someone as not having cancer when in fact they do have it.
- A good classification procedure should
  - result in few misclassifications
  - take ***prior probabilities of occurrence*** into account
  - consider the cost of misclassification
- For example, suppose there tend to be more financially sound firms than bankrupt firm. If we really believe that the prior probability of a financially distressed and ultimately bankrupted firm is very small, then one should classify a randomly selected firm as non-bankrupt unless the data overwhelmingly favor bankruptcy.
- There are two costs associated with discriminant analysis classification: The true misclassification cost per class, and the expected misclassification cost (ECM) per observation.
- **Example** Suppose there we have a binary classification system, with two classes: class 1 and class 2. Suppose that classifying a class 1 object as belonging to class 2 represents a more serious error than classifying a class 2 object as belonging to class 1. There would an assignable cost to each error. **c(i|j)** is the cost of classifying an observation into class j if its true class is i. The costs of misclassification can be defined by a cost matrix.

	Predicted Class 1	Predicted Class 2
Class 1	0	$c(2\|1)$
Class 2	$c(1\|2)$	0

Expected cost of misclassification (ECM)

- Let p_1 and p_2 be the prior probability of class 1 and class 2 respectively. Necessarily $p_1 + p_2 = 1$.
- The conditional probability of classifying an object as class 1 when it is in fact from class 2 is denoted $p(1|2)$.

- Similarly the conditional probability of classifying an object as class 2 when it is in fact from class 1 is denoted $p(2|1)$.

$$ECM = c(2|1)p(2|1)p_1 + c(1|2)p(1|2)p_2$$

(In other words: the sum of the cost of misclassification times the (joint) probability of that misclassification.)

- A reasonable classification rule should have ECM as small as possible.