Contents

1 Week 6 General Theory Topics 2

1.1 Steps in Building a Predictive Model 2

1.2 Descriptive vs Predictive Models 2

1.3 Overfitting 3

1.4 Cross-Validation and Testing 3

1.5 Cross Validation 3

1.6 Error Rates 4

1.7 Cross Validation 4

1.8 Binary Classification 7

1.9 Definitions (From Week 1) 8

1.10 Misclassification Cost 9

1 Week 6 General Theory Topics

1.1 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

1.2 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.

1.3 Overfitting

- The problem with evaluating a predictive model on training data is that it does not inform you on how well the model has generalized to new unseen data.
- A model that is selected for its accuracy on the training dataset rather than its accuracy on an unseen test dataset is very likely have lower accuracy on an unseen test dataset. The reason is that the model is not as generalized. It has specalized to the structure in the training dataset. This is called **overfitting**.

1.4 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.

1.5 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.

1.6 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.

1.7 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 over-fitting, 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**.

Type III Errors

- Type III error is related to hypotheses suggested by the data, if tested using the data set that suggested them, are likely to be accepted even when they are not true.
- This is because circular reasoning would be involved: something seems true in the limited data set, therefore we hypothesize that it is true in general, therefore we (wrongly) test it on the same limited data set, which seems to confirm that it is true.
- Generating hypotheses based on data already observed, in the absence of testing them on new data, is referred to as **Post Hoc theorizing**.
- The correct procedure is to test any hypothesis on a data set that was not used to generate the hypothesis.

1.8 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.

1.9 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:

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

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

1.10 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. $\mathbf{c(i|j)}$ 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.