Digital Media Laboratory
Sharif University of Technology

Statistical Pattern Recognition

Classification: Introduction & Quality assessment

Hamid R. Rabiee
Jafar Muhammadi, Nima Pourdamghani

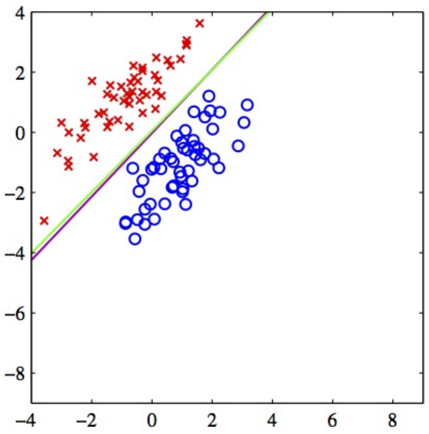
Spring 2012
http://ce.sharif.edu/courses/90-91/2/ce725-1/

Agenda

- Introduction
- Classification: A Two-Step Process
- Evaluating Classification Methods
- Classifier Performance
- Performance Measures
- Partitioning Methods

Introduction

- **Classification**
  - **predicts categorical class labels (discrete or nominal)**
  - **classifies data (constructs a model), based on the training set and the class labels, and uses it in classifying new data**



- **Typical applications**
  - **Credit approval**
  - **Target marketing**
  - **Medical diagnosis**
  - **Fraud detection**

Classification: A two-step process

- **Model construction**
  - **Each sample is assumed to belong to a predefined class, as determined by the class label**
  - **The set of samples used for model construction is called "training set"**
  - **The model is represented as classification rules, decision trees, probabilistic model, mathematical formulae and etc.**
- **Model usage**
  - **for classifying future or unknown objects**
  - **Estimate accuracy of the model**
    - The known label of test sample is compared with the classified result from the model
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model
    - Test set is independent of training set, otherwise over-fitting will occur
  - **If the accuracy is acceptable, use the model to classify data samples whose class labels are not known**

Evaluating classification methods

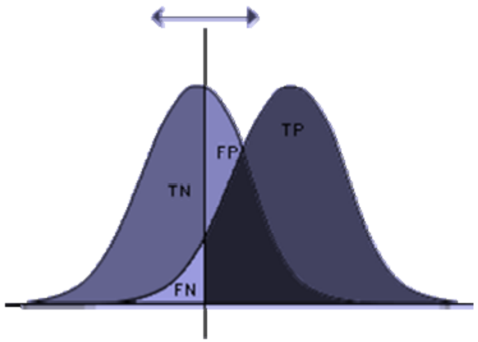
- **Performance**
  - **classifier performance: predicting class label**
    - Accuracy, {true positive, true negative}, {false positive, false negative}, …
- **Time Complexity**
  - **time to construct the model (training time)**
    - the model will be constructed once
    - can be large
  - **time to use the model (classification time)**
    - must be tolerable
    - need for good data structures
- **Robustness**
  - **handling noise and missing values**
  - **handling incorrect training data**

Evaluating classification methods

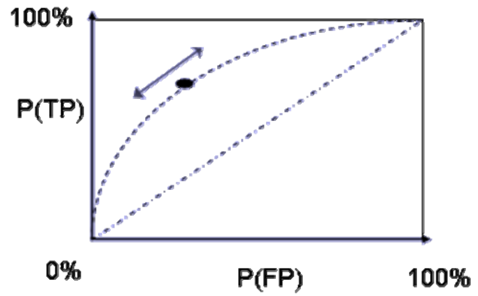
- **Scalability**
  - **efficiency in disk-resident databases**
- **Interpretability**
  - **understanding and insight provided by the model**
- **Other measures: goodness of rules or compactness of classification rules**
  - **rule of thumb: more compact, better generalization**

Performance measures

- Accuracy is not a good measure for classifier performance always (Why?)
  - Suppose a "cancer detection" problem
- Presentation of Classifier Performance
  - Use a confusion matrix or a receiver-operating characteristic (ROC) curve



Predicted \ Real	P	N
P	TP	FP
N	FN	TN



  - We can extract some performance measures from the above matrix (or curve)

Performance measures

✧ ROC Example: ROC Space

✧ **A:**

TP: 63	FP: 28	91
FN: 37	TN: 72	109
100	100	200

Acc: 0.68

✧ **B:**

TP: 77	FP: 77	154
FN: 23	TN: 23	46
100	100	200

Acc: 0.50

✧ **C:**

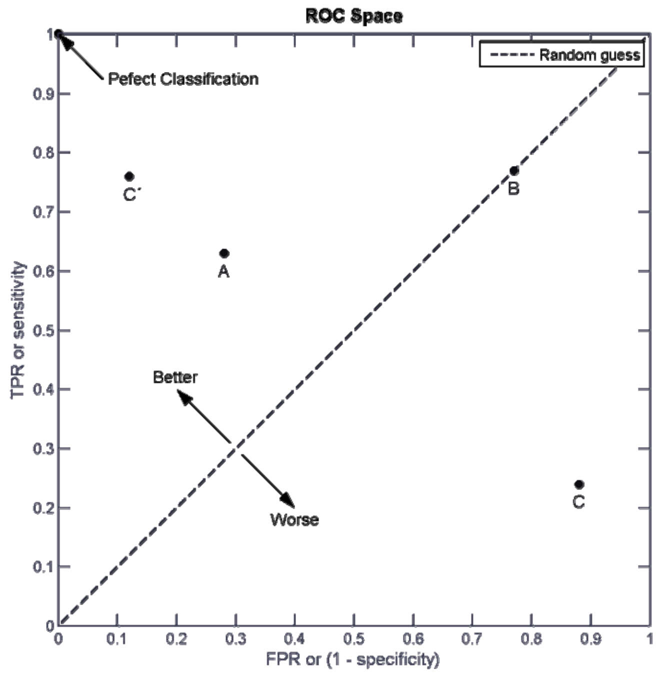
TP: 24	FP: 88	112
FN: 76	TN: 12	88
100	100	200

Acc: 0.18

✧ **C':**

TP: 76	FP: 12	88
FN: 24	TN: 88	112
100	100	200

Acc: 0.82



Performance measures

- **Performance Measures**
  - **Accuracy: (TP+TN) / (#data)**
  - **Specificity: TN / (FP+TN)**
  - **Sensitivity: TP / (FN+TP)**
  - **Index of Merit: (Specificity + Sensitivity) / 2 = (TP%+TN%) / 2**
    - Also known as “percentage correct classifications”
- **Performance measured using test set results**
  - **Test set should be distinct and different from the train (learning) set.**
  - **Several methods are available to partition the data into separated training and testing sets, resulting in different estimates of the “true” index of merit**

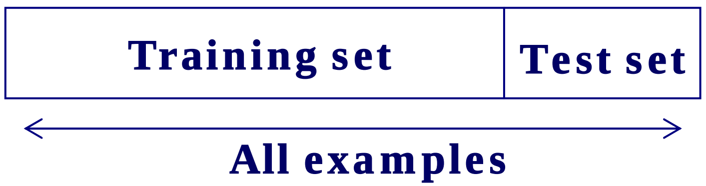
Data partitioning

- **Goal: validating the classifier and its parameters**
  - **Choose the best parameter set**
- **Idea: use a part of training data as the validation set**
- **Validation set must be a good representative for the whole data**
- **How to partition the training data**

Data partitioning methods

✧ Holdout methods: Random Sampling

- data is randomly partitioned into two independent sets
- Always size of train set is twice of test set
- Assumption: data is uniformly distributed

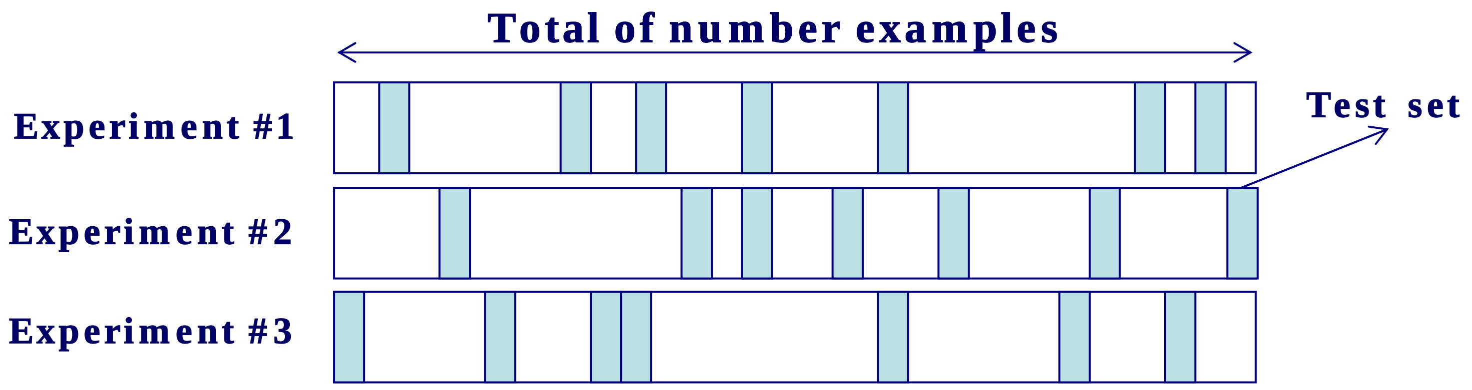


✧ Holdout methods: Bootstrap

- resample with replacement n sample of original data as training set.
- Some numbers in the original sample may be included several times in the bootstrap sample (63.2% of samples are distinct)

Data partitioning methods

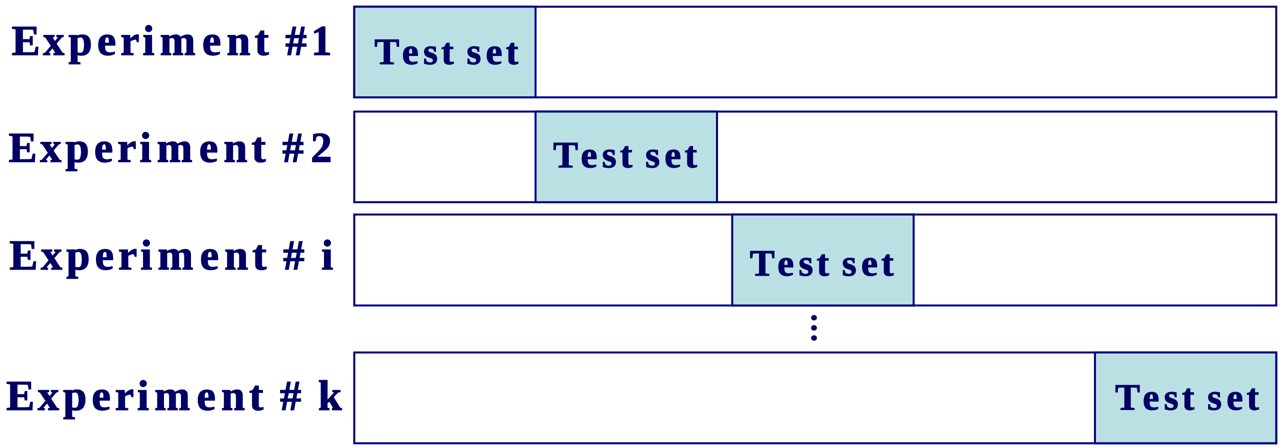
- Holdout methods: Multiple train-and-test experiment Bootstrap



- Holdout methods Drawbacks
  - In problems where we have a sparse dataset we may not be able to afford the "luxury" of setting aside a portion of the dataset for testing.
  - Since it is a single train-and-test experiment, the holdout estimate of error rate will be misleading if we happen to get an "unfortunate" split.

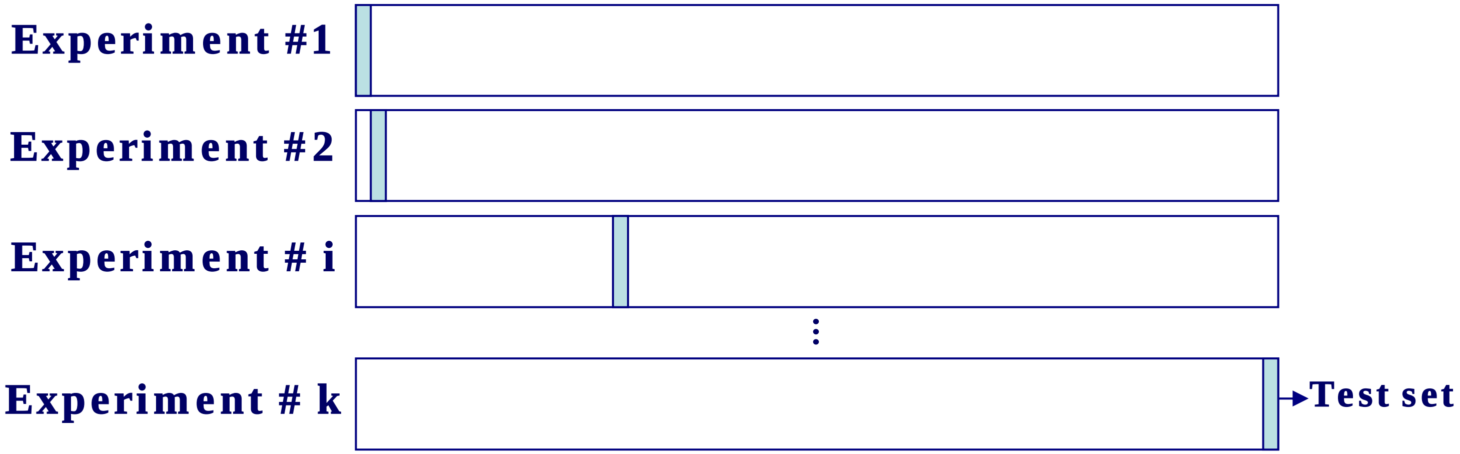
Data partitioning methods

- **Cross-validation (k-fold, where k = 10 is most popular)**
  - **Randomly partition the data into k mutually exclusive subsets, each approximately equal size**
  - **At i^{th} iteration, use D_i as test set and others as training set**
  - **The mean of measures obtained in iterations used as output performance measure**



Data partitioning methods

- **Leave-one-out**
  - **k folds where k = # of samples, for small sized data**



Data partitioning methods

- **Stratified cross-validation**
  - **folds are stratified so that class distributions in each fold is approximate the same as that in the initial data**

How many folds are needed?

- With a large number of folds
  - + The bias of the true error rate estimator will be small (the estimator will be very accurate)
  - - The variance of the true error rate estimator will be large
  - - The computational time will be very large as well (many experiments)
- With small number of folds
  - + The number of experiments and, therefore, computation time are reduced
  - + The variance of the estimator will be small
  - - The bias of the estimator will be large(conservative or higher than the true error rate)
- In practice, the choice of the number of folds depends on the size of the dataset
  - For large datasets, even 3-Fold Cross Validation will be quite accurate
  - For very sparse datasets, we may have to use leave-one-out in order to train on as many examples as possible

Three-way data splits

- **If model selection and true error estimates are to be computed simultaneously, the data needs to be divided into three disjoint sets**
  - **Training set: a set of examples used for learning: to fit the parameters of the classifier**
  - **Validation set: a set of examples used to tune the parameters of a classifier**
  - **Test set: a set of examples used <u>only</u> to assess the performance of a fully-trained classifier**
- **Why separate test and validation sets?**
  - **The error rate estimate of the final model on validation data will be biased(smaller than the true error rate) since the validation set is used to select the final model**
  - **After assessing the final model with the test set, YOU MUST NOT tune the model any further**

Any Question?

End of Lecture 5

Thank you!

Spring 2012

http://ce.sharif.edu/courses/90-91/2/ce725-1/

Template designed by Jafar Muhammadi