Data Mining & Machine Learning

Yong Zheng
Illinois Institute of Technology
Chicago, IL, 60616, USA

ILLINOIS TECH
College of Computing

Schedule

- Imbalance Issue in Classification
- Classification: Summary
- Data Preprocessing: Feature Selection & Reduction

Schedule

- Imbalance Issue in Classification
- Classification: Summary
- Data Preprocessing: Feature Selection & Reduction

Imbalance Issue

- Imbalance issue: Example
  - 100 examples, 60 are positive, 40 are negative
  - 100 examples, 90 are positive, 10 are negative
- Solutions: assume we have more positive samples
  - Undersampling [lose information, final data is small]
    Remove some positive samples in the training
    Try to obtain a balance between positives & negatives
  - Oversampling [may result in overfitting]
    Replicate and add more negative data into training
    Try to obtain a balance between positives & negatives

Imbalance Solutions: Examples

- Imbalance issue: Example
  - 100 examples, 95 are positives, 5 are negatives
  - Data split: training 83, testing 17
    In training set, 80 are positives, 3 are negatives
- Solutions: [applied to training set only]
  - Undersampling [lose information, final data is small]
    Use only 3 positives & 3 negatives in training set
  - Oversampling [may result in overfitting]
    Use 80 positives and 80 negatives in training set
    Replicate the 3 negatives to have 80 negatives

Schedule

- Imbalance Issue in Classification
- Classification: Summary
- Data Preprocessing: Feature Selection & Reduction

Classification: Summary

- Classification Tasks
  - Binary Classification
  - Multi-Class Classification
    - Can be solved by algorithms directly, e.g., KNN, Trees, etc.
    - Can be solved by multiple binary classification, e.g., one vs rest
  - Multi-Label Classification
    - Can be solved by several binary/multi-class classifications
    - Can be solved by well-designed MLC models
    - Have special evaluation metrics, rather than the traditional metrics in classifications

Classification: Summary

- Classification Algorithms
  - KNN, Naïve Bayes, Decision Tree, Logistic Regression, SVM, Neural networks (will discuss it in future)
  - Ensemble methods: they are just ensembling methods, not classification algorithms. They can work together with any classification algorithms
  - Knowledge points
    - Understand how each method works
    - Know the requirements (e.g., data) to use these methods
    - Know the key parameters to be tuned up
    - Know how to alleviate overfittings and imbalance issues

Classification: Summary

- Evaluation Metrics
  - Accuracy
  - Precision
  - Recall
  - F1
  - ROC Curve
  - Showing overall accuracy only is not enough to evaluate a classification model

Schedule

- Imbalance Issue in Classification
- Classification: Summary
- Data Preprocessing: Feature Selection & Reduction

Feature Selection and Reduction

- This is very important process in data analytics and data mining.
- Reason why?
  - Not all of the features are useful
  - Irrelevant features will decrease accuracy
  - Data collection is an expensive process, you cannot simply remove features with your common sense
  - You must remove features or reduce dimensions by specific reasons

Major Techniques of Dimensionality Reduction

- Feature selection
  - Definition
  - Objectives

- Feature Extraction (reduction)
  - Definition
  - Objectives

Feature Selection

- Definition
  - A process that chooses an optimal subset of features according to a objective function
- Objectives
  - To reduce dimensionality and remove noise
  - To improve mining performance
    - Speed of learning
    - Predictive accuracy
    - Simplicity and comprehensibility of mined results

Feature Extraction/Reduction

- Feature reduction refers to the mapping of the original high-dimensional data onto a lower-dimensional space
- Given a set of data points of p variables $\{x_1, x_2, \cdots, x_n\}$

Compute their low-dimensional representation:

$$x_i \in \Re^d \rightarrow y_i \in \Re^p \ (p << d)$$

- Criterion for feature reduction can be different based on different problem settings.
  - Unsupervised setting: minimize the information loss, e.g., PCA
  - Supervised setting: maximize the class discrimination, e.g., LDA

Feature Reduction vs. Feature Selection

- Feature reduction
  - Input: All original features are used
  - Output: The transformed features are linear combinations of the original features
- Feature selection
  - Output: Only a subset of the original features are selected

Feature Reduction vs. Feature Selection

- Feature Selection
  - Filtering approach — Kohavi and John, 1996
  - Wrapper approach — Kohavi and John, 1996
  - Embedded methods — I. Guyon et. al., 2006

- Dimensionality Reduction
  - Principal Components Analysis (PCA) — http://www.cs.otago.ac.nz/cosc453/student_tutorials/...principal_components.pdf
  - Nonlinear PCA (Kernel PCA, CatPCA) — Schoelkopf et. al., 2001; .;Gifi, 1990
  - Multi-Dimensional Scaling (MDS) — Born and Groenen, 2005
  - Homogeneity Analysis — Gifi, 1990

Feature Selection

Components In Feature Selection

- For every feature selection technique, there must be at least two components
  - Quality Measure
  - Search/Rank Methods

Example: Linear Regression

- In linear regression, we are going to predict a numerical variable y, by using a set of x variables, e.g., $x_1, x_2, x_3, \ldots, x_n$
- Search Methods
  - Backward Elimination
    Use all x variables to build the model
    Drop x variables step by step to see whether we can improve the model
  - Forward Selection
    Build a simple model, e.g., a model with only one x
    Try to add more x variables step by step to see whether we can improve the model
  - Stepwise = Forward + Backward

Example: Linear Regression

- In linear regression (introduced in ITMD 527), we discuss different ways to select independent variables to predict the dependent variable
  - Backward Elimination by using p-value
  - Backward Elimination by using AIC/BIC
  - Forward Selection or Stepwise by using AIC/BIC

Search or Rank Method

Quality Measures

Quality Measures

- The goodness of a feature/feature subset is dependent on measures
- Various measures
  - Information measures (Yu & Liu 2004, Jebara & Jaakkola 2000)
  - Distance measures (Robnik & Kononenko 03, Pudil & Novovicov 98)
  - Dependence measures (Hall 2000, Modrzejewski 1993)
  - Consistency measures (Almuallim & Dietterich 94, Dash & Liu 03)
  - Accuracy measures (Dash & Liu 2000, Kohavi&John 1997)

Information Measures

- Entropy of variable *X*

Impurity Measure

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i))$$

- Entropy of *X* after observing *Y*

$$H(X|Y) = -\sum_j P(y_j)\sum_i P(x_i|y_j)\log_2(P(x_i|y_j))$$

- Information Gain

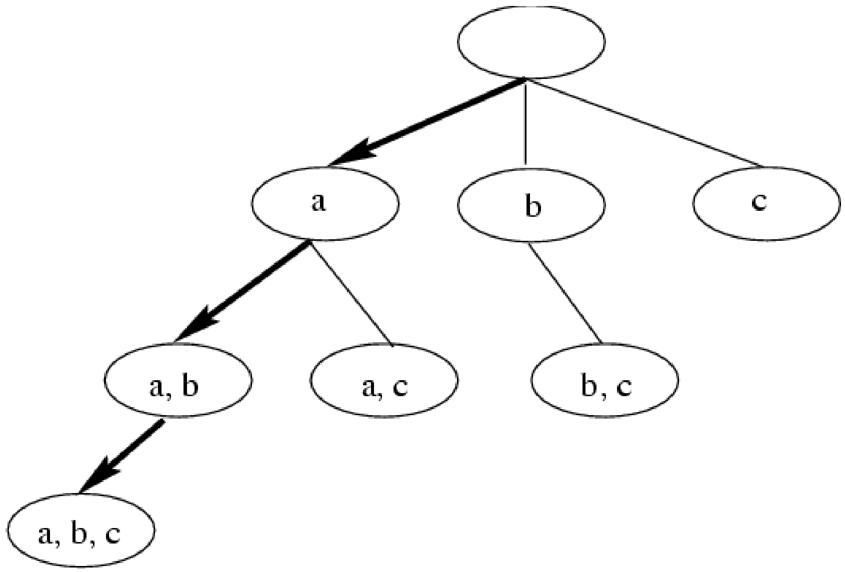
This measure is used in decision tree classification

$$IG(X|Y) = H(X) - H(X|Y)$$

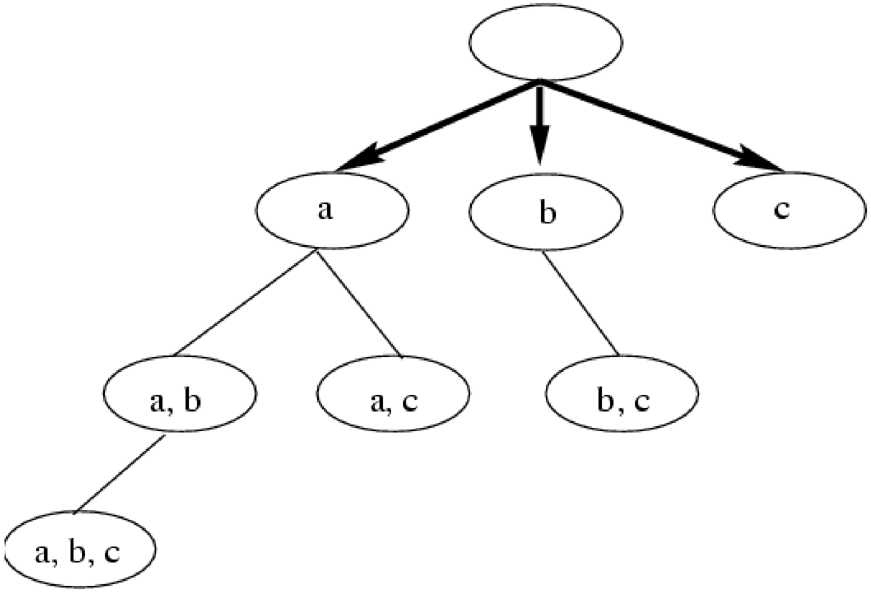
Accuracy Measures

- Using classification accuracy of a classifier as an evaluation measure
- Factors constraining the choice of measures
  - Classifier being used
  - The speed of building the classifier
- Compared with previous measures
  - Directly aimed to improve accuracy
  - Biased toward the classifier being used
  - More time consuming

Feature Search



Depth-first search



Breadth-first search

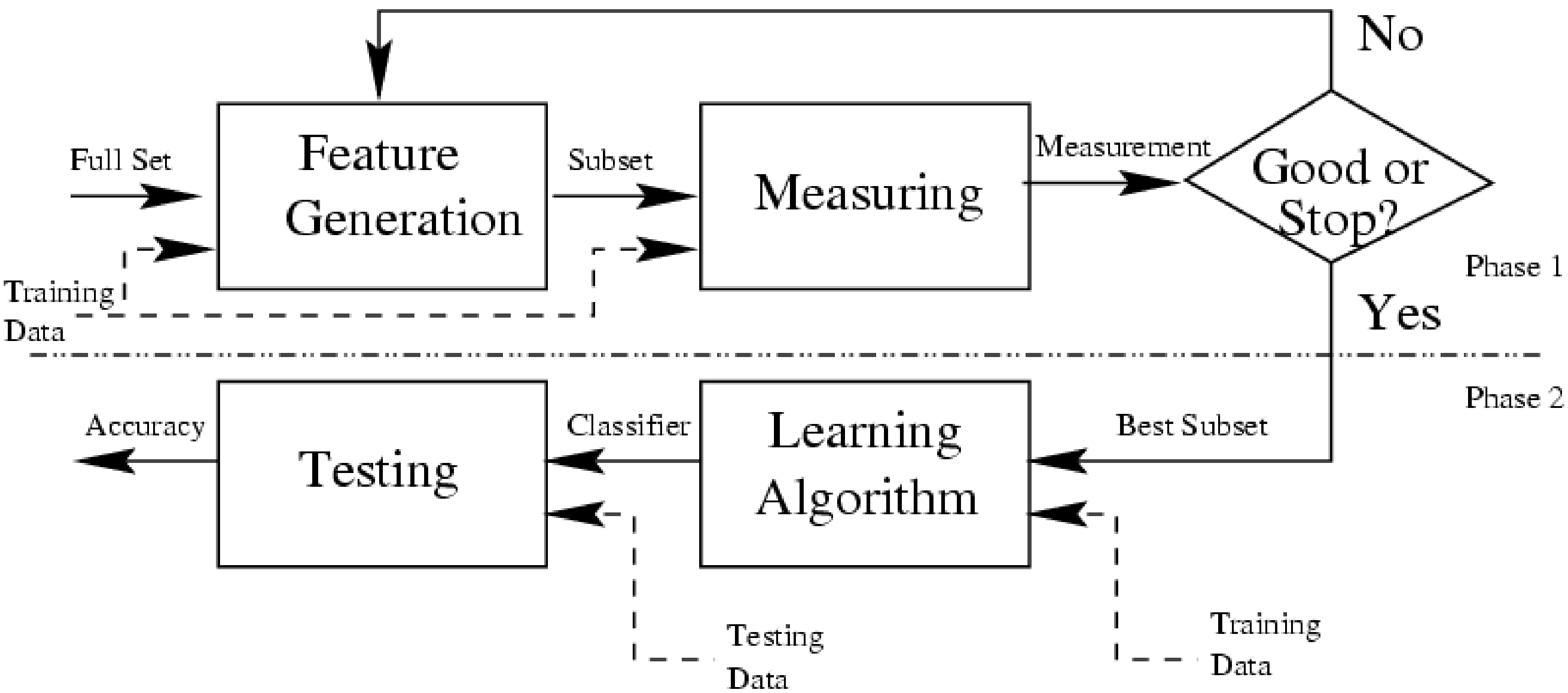
Feature Ranking

- Weighting and ranking individual features
- Selecting top-ranked ones for feature selection
- Advantages
  - Efficient: $O(N)$ in terms of dimensionality N
  - Easy to implement
- Disadvantages
  - Hard to determine the threshold
  - Unable to consider correlation between features

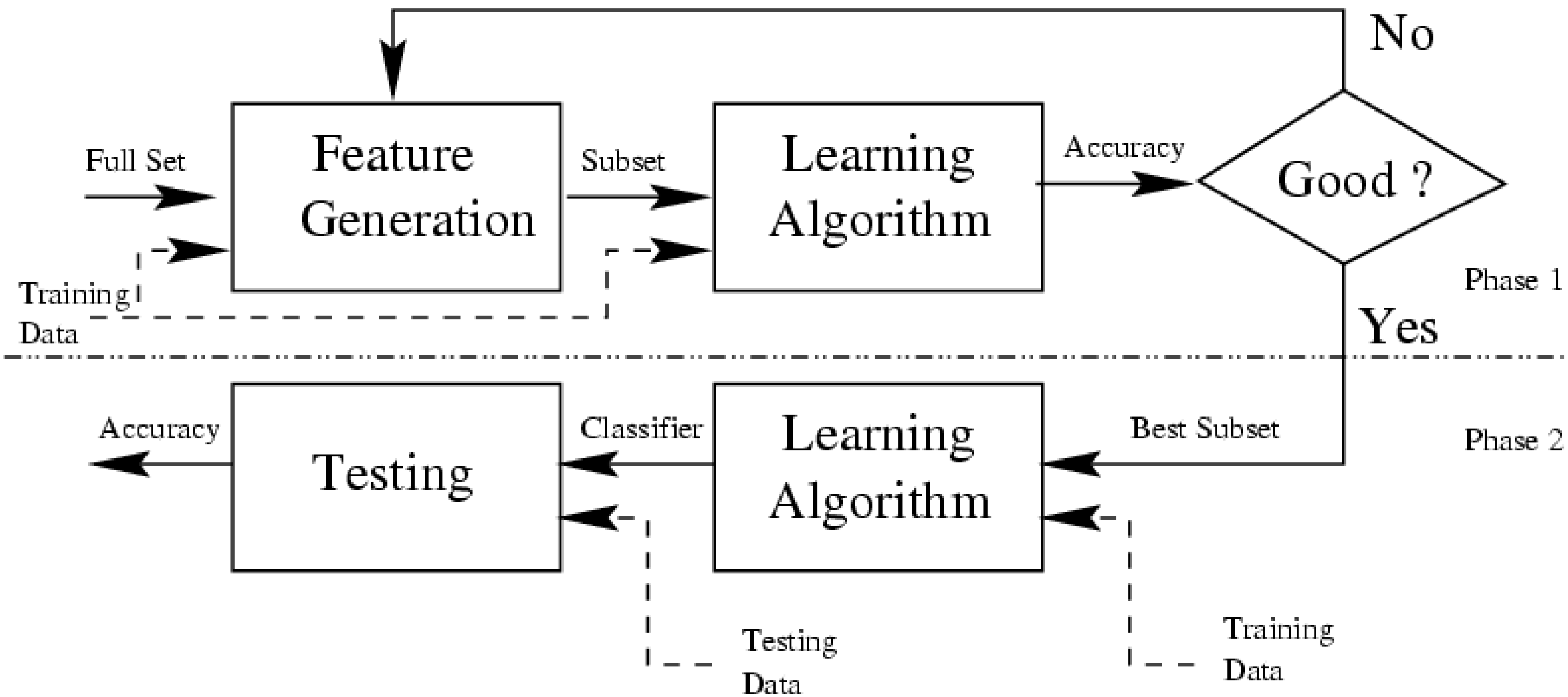
Two Models of Feature Selection

- Filter model
  - Separating feature selection from classifier learning
  - Relying on general characteristics of data (*information, distance*, *dependence*, *consistency*)
  - No bias toward any learning algorithm, fast running
- Wrapper model
  - Relying on a pre-determined classification algorithm
  - Using predictive accuracy as goodness measure
  - High accuracy, computationally expensive

Filter Model



Wrapper Model



Feature Reduction

Feature Reduction Algorithms

- Unsupervised
  - Latent Semantic Indexing (LSI): truncated SVD
  - Independent Component Analysis (ICA)
  - Principal Component Analysis (PCA)
  - Manifold learning algorithms
- Supervised
  - Linear Discriminant Analysis (LDA)
  - Canonical Correlation Analysis (CCA)
  - Partial Least Squares (PLS)
- Semi-supervised

Feature Reduction Algorithms

Linear Discriminant Analysis (LDA) tries to identify attributes that account for the most variance between classes. In particular, LDA, in contrast to PCA, is a supervised method, using known class labels.

Principal Component Analysis (PCA) applied to this data identifies the combination of linearly uncorrelated attributes (principal components, or directions in the feature space) that account for the most variance in the data. Here we plot the different samples on the 2 first principal components.

Singular Value Decomposition (SVD) is a factorization of a real or complex matrix. Actually SVD was derived from PCA.

Principal Component Analysis

Assume we have a data with multiple features

1). Try to find principle components (PCs) – each component is a combination of the linearly uncorrelated attributes/features;

2). PCA allows to obtain an ordered list of those components that account for the largest amount of the variance from the data;

3). The amount of variance captured by the first component is larger than the amount of variance on the second component, and so on.

4). Then, we can reduce the dimensionality by ignoring the components with smaller contributions to the variance.

5). The final reduced features we have are no longer the original features, but the difference PCs, each PC is a linear combination of your original features.

Principal Component Analysis

How to obtain those principal components?

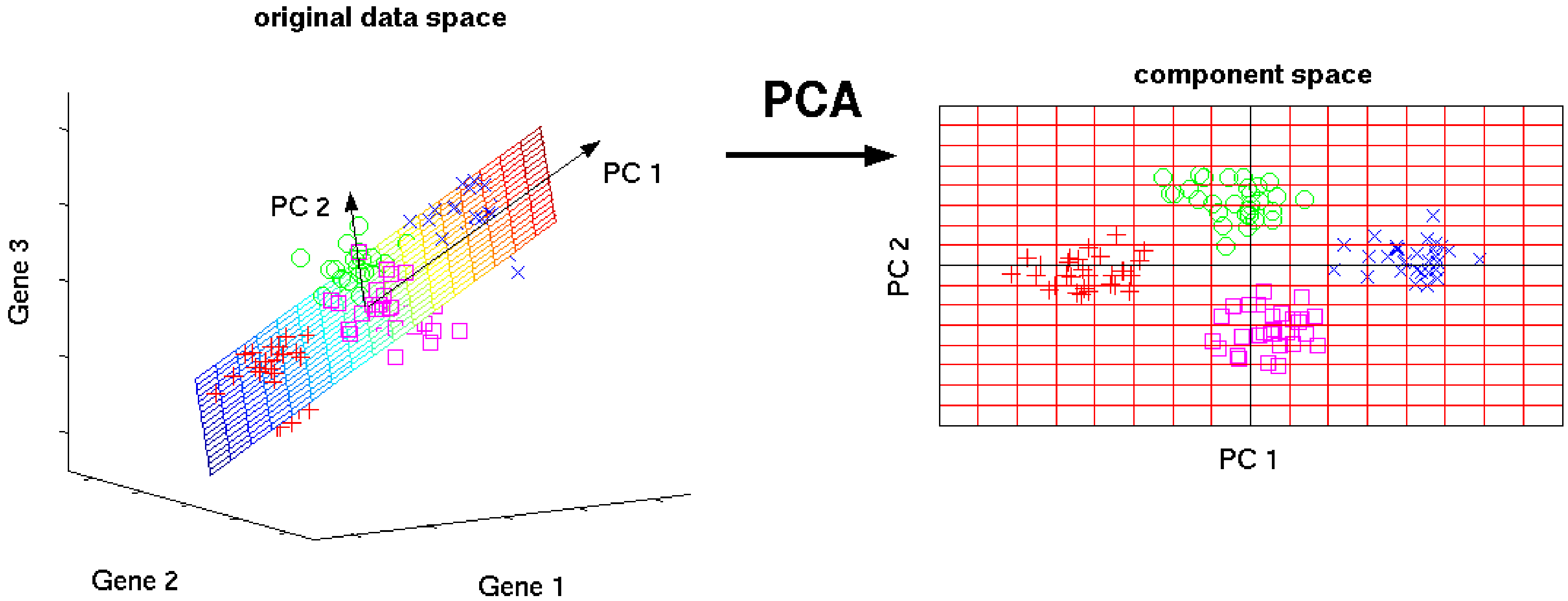
The basic principle or assumption in PCA is:
The eigenvector of a covariance matrix equal to a principal component, because the eigenvector with the largest eigenvalue is the direction along which the data set has the maximum variance.

Each eigenvector is associated with a eigenvalue;
Eigenvalue ➔ tells how much the variance is;
Eigenvector ➔ tells the direction of the variation;

The next step: how to get the covariance matrix and how to calculate the eigenvectors/eigenvalues?

Visualization of PCA

Example: Gene Expression



The original expression by 3 genres is projected to two new dimensions, Such two-dimensional visualization of the samples allow us to draw qualitative conclusions about the separability of experimental conditions (marked by different colors).

Feature Reduction vs. Feature Selection

- Feature reduction
  - Input: All original features are used
  - Output: The <u>transformed features</u> are linear combinations of the original features
- Feature selection
  - Output: Only <u>a subset of the original features</u> are selected