Data Mining - Lecture Notes

Maastricht University

October 16, 2025

Contents

Data Mining - Lecture Notes

Maastricht University

October 16, 2025

Contents

1 Getting to Know the Data

Lecture 2: Getting to Know the Data

Data

Data objects are described by variables. A variable V represents a property or characteristic of an object that may vary, either from one object to another or from one time to another.

Variable Definition

A variable V is a quadruple $\langle$Name, Domain, Operations, Scale$\rangle$:

- **Name**: The name of V
- **Domain**: The set of values of V
- **Operations**: The set of operations allowed over Domain
- **Scale**: A rule that associates a value from Domain for the variable V when it represents an object o.

Types of Data

- Record Based Data (Transactions)
- Graph Based Data (WWW)
- Ordered Data (Genomics)

Measuring Data

Mean

- Population Mean: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
- Sample Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Weighted Mean: $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

Median

Middle value if odd number of values, average of middle two values if even number of values

Mode

Value that occurs most frequently

Symmetric vs Skewed Data

- **Symmetric**: Data is symmetrically distributed around the mean. The mean, median, and mode are all equal.
- **Skewed**: Data is not symmetrically distributed around the mean.
  - **Left Skewed (Negative)**: Mean < Median < Mode
  - **Right Skewed (Positive)**: Mode < Median < Mean

Dispersion

- **Range**: $\max(x) - \min(x)$
- **Quantile**: At most $n(k/q)$ values will be smaller
- **IQR**: $Q_3 - Q_1$ (middle 50% of data)
- **Outliers**: Values outside $[Q_1 - 1.5 \times \text{IQR}, Q_3 + 1.5 \times \text{IQR}]$
- **Variance**: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
- **Standard Deviation**: $\sigma = \sqrt{\sigma^2}$

Data Characteristics

- **Dimensionality**: Number of variables/features. The curse of dimensionality refers to the exponential increase in data required to densely populate space as the dimension increases.
- **Sparsity**: Proportion of missing/zero values in the data.
- **Resolution**: Level of detail or aggregation in the data.

Normal Distribution

- Bell-shaped curve
- $\mu = \bar{x}$
- $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$
- **68-95-99.7 Rule**:
  - 68% within $\mu \pm \sigma$
  - 95% within $\mu \pm 2\sigma$
  - 99.7% within $\mu \pm 3\sigma$

Variable Types

- **Nominal**: Categories without order (gender, color, zip code)
- **Ordinal**: Ordered categories (education level, income level)
- **Interval**: Ordered with equal intervals (temperature in °C, dates)
- **Ratio**: Interval with true zero (height, weight, age)

Statistical Plots

- **Boxplot**: Five-number summary (min, Q1, median, Q3, max) and outliers
- **Histogram**: Shows frequency distribution of numerical data
- **Quantile Plot**: Plots data against theoretical quantiles (index $f = \frac{i-0.5}{n}$)
- **Q-Q Plot**: Compares two distributions using their quantiles
- **Scatter Plot**: Shows relationship between two numerical variables

Outlier Handling

1. Remove if erroneous
2. Transform (log, square root)
3. Use robust statistics (median, IQR)
4. Cap/floor extreme values

2 Data Preprocessing

Lecture 3: Data Preprocessing

Overview

The detection and correction of data quality problems. The use of algorithms that can tolerate poor data quality. Data can be:

- Inconsistent: data transformations, technology problems, human errors
- Incomplete: missing values, incomplete records
- Inaccurate: errors in data entry, data transformations, technology problems
- Outdated: data transformations, technology problems

Data Cleaning

Converting data so that it becomes consistent, complete, accurate, and up-to-date. It's realized by filling missing values, removing duplicates, smoothing noise, and resolving inconsistencies.

Handling Noisy Data

- Clustering (detect and remove outliers)
- Computer and human inspection

Handling Missing Data

- Filling manually
- Using the variable mode, median, or mean

Data Integration

Combining data from different sources.

Possible Problems

- Different variables have the same name
- Similar variables have different names
- Redundant variable: can be detected with Chi-Square, Covariance analysis

Chi-square Test

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Covariance

$$\text{Cov}(X, Y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)$$

Data Reduction

Obtaining a reduced set of variables that are sufficient to represent the data.

Strategies

- **Principal Component Analysis (PCA)**: Converts variables into a new set of variables that are uncorrelated and capture the maximum variance.
- **Multidimensional Scaling (MDS)**: Finds a low-dimensional representation that preserves pairwise distances (or dissimilarities) between points.
- **Feature Selection**: Select a subset of variables that are most relevant to the task.
- **Clustering**: Group similar objects together.
- **Sampling**: Main strategy for data reduction in data mining. The sample must be representative of the population.
  - Without replacement: each object is selected only once
  - With replacement: each object can be selected multiple times
  - Stratified sampling: data is split into partitions and a sample is taken from each partition

Data Valuation

Seeks to assign a numerical value to an individual's data in the trade of data. The issue is the time and cost of data valuation. Complexity is above $O(2^N)$.

Data Transformation and Discretization

A function that maps the entire set of values of a given variable to a new set of replacement values.

Methods

- **Normalization**: Scales the values to a range, such as [0, 1] or [-1, 1].
- **Smoothing**: Reduces noise in the data.
- **Variable/Feature Construction**: Creates new variables from existing variables.

Normalization

- **Min-Max**: Scales the values to a range, such as [0, 1] or [New Min, New Max]

$$\bar{x} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- **Z-Score**: Scales the values to have a mean of 0 and a standard deviation of 1.

$$\bar{x} = \frac{x - \mu}{\sigma}$$

Discretization

Divides the range of continuous values into a set of intervals. The intervals are called bins and can replace the original values. Clustering can also be used to find the intervals.

- **Binning**: Can be done with equal width or equal frequency (depth)

3 Regression

Lecture 4: Regression

Regression

Given a set of variables X, we want to predict a target variable Y. There's an unknown function f that maps X to Y.

We assume that $Y = f(X) + \epsilon$ where ϵ is the noise. $f(X)$ is an optimal function that minimizes the error. The random error term ϵ is assumed to be independent of X and has a mean of 0 and cannot be reduced by any model.

The optimal function $f(X)$ is the one that minimizes the error. The error is defined as the difference between the predicted value and the actual value.

Parametric Methods

Parametric methods assume a specific form for the function $f(X)$. They involve two steps:

1. We assume a form for $f(X)$.
$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$
2. Using training data, we find the parameters of $f(X)$ that minimize the error.

Because of the restricted flexibility of parametric methods, they are open to **underfitting**.

- **Underfitting**: The model is too simple to capture the relationship between X and Y.

Non-Parametric Methods

Non-parametric methods do not assume a specific form for the function $f(X)$. They are flexible and can capture complex relationships between X and Y, but they require more data to estimate the function. They are open to **overfitting**.

- **Overfitting**: The model corresponds too closely to the training data and fails to generalize to new data.

Assessing Model Performance

- **Mean Squared Error (MSE)**:
$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Bias-Variance Tradeoff

The expected error of a model is the sum of three components:

- **Bias**: The error due to the model's inability to capture the relationship between X and Y.
- **Variance**: The error due to the model's sensitivity to the training data.
- **Irreducible Error**: The error due to the noise in the data.

$$\text{Expected Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

Linear Regression

The linear regression model assumes a linear relationship between X and Y. We assume a true dependency:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

Assessing the Model

- **Residuals**: $y_i - \hat{y}_i$
- **Residual Sum of Squares (RSS)**: $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
- **Residual Standard Error (RSE)**: $\sqrt{\frac{\text{RSS}}{N-2}}$
- **R-squared**: $R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$
- **Total Sum of Squares (TSS)**: $\text{TSS} = \sum_{i=1}^{N}(y_i - \bar{y})^2$

RSE is the standard deviation of the residuals, or the irreducible error epsilon.

Additive Assumption

The effect of any variable is independent of the values of the effects of other variables.

To avoid the additive assumption, we can use interaction terms. We move from the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

To the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \epsilon$$

The interaction term introduces a non-linear relationship between X_1 and X_2.

Shrinkage Methods

Shrinkage methods reduce the variance of the model by shrinking the coefficients of the variables.

Ridge Regression

Ridge regression adds a penalty term to the RSS:

$$\text{RSS} + \lambda \sum_{i=1}^{p} \beta_i^2$$

Where λ is the regularization parameter.

The Ridge Regression model is:

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y$$

Lasso Regression

Lasso regression adds a penalty term to the RSS:

$$\text{RSS} + \lambda \sum_{i=1}^{p} |\beta_i|$$

This has the effect of forcing some of the coefficients to zero. So the Lasso performs feature selection.

KNN Regression

KNN regression is a non-parametric method that uses the k-nearest neighbors to predict the value of a new point.

The KNN model is:

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} y_i$$

The value of k is a hyperparameter that needs to be tuned. If k is small, then the regression model is more flexible.

However, this regression is not so good in high dimensions.

Decision Tree Regression

The decision tree regression model is a non-parametric method that uses a tree to predict the value of a new point.

Regression tree is learned by minimizing the RSS:

$$\text{RSS} = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_j)^2$$

Where R_j is the number of regions in the tree.

Regression Trees are sensitive to overfitting. To avoid this, we can use pruning. There are two approaches:

1. **Pre-pruning**: Stop the tree before it overfits.
2. **Post-pruning**: Prune the tree after it overfits.

Model Trees provide different values for different instances due to the regression models in the leaves.

4 Classification

Lecture 5: Classification

Classification

Naive Bayes

This is a parametric method that uses Bayes' theorem to predict the class of a new point.

It is based on the Bayes' theorem:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

It assumes that the features are independent given the class and are identically distributed. Independent samples are samples that are drawn from the same distribution.

There are two types of classifiers:

- **Discrete classifiers**: assign a class label to a test instance
- **Score classifiers**: assign a continuous score for each class and can be assigned to a test instance.

An optimal Bayes rule assumes knowledge of:

- The prior distribution p(y)
- The distribution p(x—y) for each class y
- The distribution p(x)

Naïve Bayes classifier naively assumes that the input variables are conditionally independent. The function is:

Naïve Bayes can be a linear or non-linear method for classification depending on the properties of the input variables.

Naïve Bayes does not have any parameter to control the bias-variance trade-off. The only way can be explicit feature selection.

Logistic Regression

Instead of predicting Y, we predict $P(Y = 1|X)$ using the logistic (sigmoid) function:

$$P(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p}}$$

For more than two classes we use the softmax function:

$$P(Y = k|X) = \frac{e^{\beta_{k0}+\beta_{k1} X_1+\beta_{k2} X_2+\ldots+\beta_{kp} X_p}}{\sum_{i=1}^{K} e^{\beta_{i0}+\beta_{i1} X_1+\beta_{i2} X_2+\ldots+\beta_{ip} X_p}}$$

We can estimate the parameter β using maximum the likelihood function:

$$l(\beta_0, \beta) = \prod_{i=1}^{n} P(Y = y_i|x_i)$$

Where if $Y = 1$:

$$P(Y = y_i|x_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}}}{1 + e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}}}$$

And if $Y = 0$:

$$P(Y = y_i|x_i) = \frac{1}{1 + e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}}}$$

Summary

- Logistic regression is a parametric method for classification.
- Logistic regression is a linear method for classification.
- Logistic regression estimates class probabilities. It does not make a classification decision; i.e., it is a scoring classifier.
- The variance of logistic regression can be reduced using shrinkage methods based on ridge regression (ridge logistic regression).

Support Vector Machine

Support Vector Machines (SVMs) approach the two-class classification problem in a direct way. SVM tries indeed to separate the classes in instance space X.

Among all separating hyperplanes, find the one that makes the biggest gap or margin between the two classes.

To convert SVM to a scoring classifier, we can use the decision function:

$$f(x) = \frac{1}{1 + e^{Af(x)+B}}$$

where A and B are parameters that can be estimated using maximum likelihood.

Higher C values imply low flexibility (high bias, low variance). Lower C values imply high flexibility (low bias, high variance).

Summary

- SVM is a discrete classifier. It provides a classification (no probability)!
- SVM can be converted to a scoring classifier using signed distance to hyperplane (directly or using the Platt scaling).
- SVM is a parametric method for binary classification.
- SVM is a linear method for classification.
- SVMs handle nonseparability problems using: Soft-margins and Kernels.

Feature Expansion Enlarge the feature space X by adding new features: X^2, X^3, X_1X_2, ... Fit a support-vector classifier in the enlarged space. This results in non-linear decision boundaries in the original space.

Kernel Support Vector Machine We have a kernel function $K(x_i, x_j)$ that computes the similarity between x_i and x_j. Then the decision function is:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i K(x_i, x)$$

Decision Trees

Each interior node tests a variable. Each branch corresponds to a variable value. Each leaf node is labeled with a class (class node).

```
function Classify(x: instance, node: variable containing a node of DT)
    if node is a classification node then
        return the class of node
    else
        determine the child of node that matches x
        return Classify(x, child)
    end if
end function
```

It is okay for the training data to contain missing values. Decision trees can be used even if instances have missing variables.

Basic Algorithm:

1. X ← the ”best” decision variable for a node N.
2. Assign X as decision variable for the node N.
3. For each value of X, create new descendant of the node N.
4. Sort training instances to leaf nodes.
5. IF training examples perfectly classified, THEN STOP. ELSE iterate over new leaf nodes.

Entropy Let S be a sample of training examples, and p_+ is the proportion of positive examples in S and p_- is the proportion of negative examples in S. Then: entropy measures the impurity of S

$$H(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

Bias-Variance Tradeoff Decision trees have in general high variance.

- The bias of decision trees decreases with the size of the trees.
- The variance of decision trees increases with the size of the trees.

Overfitting

- **Pre-pruning**: stop growing the tree earlier, before it reaches the point where it perfectly classifies the training data.
- **Post-pruning**: Allow the tree to overfit the data, and then post-prune the tree.

Validation set is a set of instances used to evaluate the utility of nodes in decision trees. The validation set has to be chosen so that it is unlikely to suffer from same errors or fluctuations as the training set.

Summary

- DTs are discrete classifiers. They can estimate probabilities by normalizing class scores in each leaf node.
- Decision Trees (DT) for a non-parametric method for classification.
- DTs are a non-linear method for classification.

Decision Rules

Decision rules are rules with the following form:

```
if {conditions} then concept C
```

Summary

- Decision Rules (DRs) are discrete classifiers. They can estimate probabilities by normalizing class scores in each rule.
- Decision Rules form a non-parametric method for classification.
- DRs are a non-linear method for classification.
- DRs are usually simpler than decision trees on the same data.

K-NN Classification

k-NN Classifier is a non-parametric classifier. To estimate a class value y for a given test instance x, find a set NN of the k closest instances to x in training data Tr.

- **Discrete Classification**: output the majority class among the instances in NN.
- **Scoring Classification**: output the score for each class among the instances in NN. If the scores are normalized we estimate class probabilities.

The value of k controls the flexibility of the k-NN classifier. The smaller that value the more flexible is the k-NN classifier (the higher the variance and lower the bias).

Notes:

- Continuous variables should be normalized. Otherwise, the variables with bigger domains prevail!
- Discrete variables do not pose problems since distances are based on value matches.

Advantages

1. The NN classifier can estimate complex class borders locally and differently for each new test instance.
2. The NN classifier provides good generalization performance on many domains.
3. The NN classifier learns very quickly.
4. The NN classifier is robust to noisy training data.
5. The NN classifier is intuitive and easy to understand which facilitates implementation and modification.

Disadvantages

1. The NN classifier has large storage requirements because it has to store all the data.
2. The NN classifier is slow during instance classification because all the training instances have to be visited.
3. The generalization performance of the NN classifier degrades with increase of noise in the training data.
4. The generalization performance of the NN classifier degrades with increase of irrelevant variables.

Summary

- Nearest-Neighbor (NN) Classifier is a non-parametric method for classification.
- NN is a non-linear method for classification.
- NN can be a discrete classifier and a scoring classifier depending on how we handle the class statistics of the nearest neighbors.
- The bias-variance trade-off can be controlled by the parameter k.

5 Feature Selection

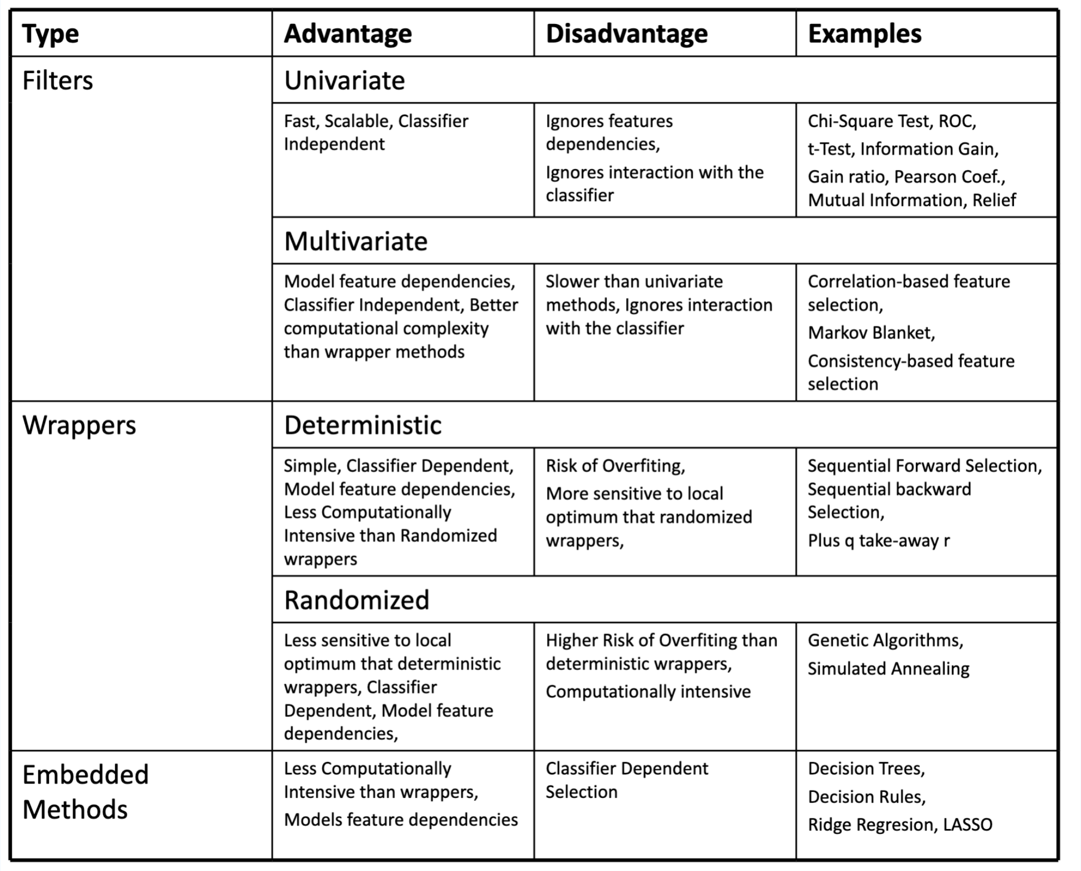
Lecture 6: Feature Selection

Feature Selection

Given a supervised learning task, feature selection is a process of selecting features s.t. the generalization performance of the predictive models is improved.

Objectives:

- Avoid overfitting and improve generalization performance;
- Provide faster and more cost-effective models;
- Gain a deeper insight into the underlying processes that generated the data.



Type	Advantage	Disadvantage	Examples
Filters	Univariate		
	Fast, Scalable, Classifier Independent	Ignores features dependencies, Ignores interaction with the classifier	Chi-Square Test, ROC, t-Test, Information Gain, Gain ratio, Pearson Coef., Mutual Information, Relief
	Multivariate		
	Model feature dependencies, Classifier Independent, Better computational complexity than wrapper methods	Slower than univariate methods, Ignores interaction with the classifier	Correlation-based feature selection, Markov Blanket, Consistency-based feature selection
Wrappers	Deterministic		
	Simple, Classifier Dependent, Model feature dependencies, Less Computationally Intensive than Randomized wrappers	Risk of Overfiting, More sensitive to local optimum that randomized wrappers,	Sequential Forward Selection, Sequential backward Selection, Plus q take-away r
	Randomized		
	Less sensitive to local optimum that deterministic wrappers, Classifier Dependent, Model feature dependencies,	Higher Risk of Overfiting than deterministic wrappers, Computationally intensive	Genetic Algorithms, Simulated Annealing
Embedded Methods	Less Computationally Intensive than wrappers, Models feature dependencies	Classifier Dependent Selection	Decision Trees, Decision Rules, Ridge Regresion, LASSO

Filters

Rank features or feature subsets independently of the learning algorithm (classifier).

- **Advantages:** Fast, Simple, Interpretable, Classifier independence, High dimensionality tolerance
- **Disadvantages:** Ignore interaction with classifiers, ignore feature dependencies.

Univariate Filters

The relevance of individual feature X_i is determined by how much X_i can explain the output feature Y. Statistically this means that we need to assess the (in)dependence of X_i and Y. In this context we have two questions:

1. Is X_i independent of Y? 2. If X_i and Y are not independent, then how much are they dependent?

We have these scenarios:

- The feature X_i is discrete and the feature Y is discrete: Chi-Square Test for Independence, ROC, Information gain and Ratio; Relief; Mutual Information.
- The feature X_i is continuous and the feature Y is discrete: t-Test on two means; Relief.
- The feature X_i is continuous and the feature Y is continuous: Mutual Information.

Multivariate Filters

Multivariate Filters rank feature subsets independently on the type of the predictor later used. They operate by searching in the space of possible feature subsets and choosing those of subsets that maximize a given evaluation criterion.

N features, 2^N possible feature subsets

Search

- **Search Direction:** Forward, Backward, Bidirectional
- **Search Strategy:**
  - Deterministic: Hill Climbing, Best First, Exhaustive
  - Non-deterministic: Genetic

Feature Subset Assessment

Split the data into 3 subsets: training, validation, and test.

1. For each feature subset, train predictor on training data.
2. Select the feature subset which performs best on validation data.
3. Repeat and average if you want to reduce variance (cross-validation).
4. Test on test data.

Data Leakage

Data leakage in validation occurs when information from outside the training fold is used—directly or indirectly—during model training or preprocessing, causing the model to appear more accurate in evaluation than it really is.

Correlation Based Feature Selection

Evaluation of feature subsets based on the next formula:

$$\mathrm{CFS}(X_i, Y) = \frac{k r_{cf}}{\sqrt{k + (k-1) r_{ff}}}$$

where S is a set with k features, r_{cf} is the average correlation between the features in S and the class feature, and r_{ff} is the average correlation between the features in S.

Wrappers

Wrappers rank feature subsets w.r.t. the predictor used. They operate by searching in the space of possible feature subsets and choosing those of subsets that maximize a given evaluation criterion based on the predictor later used. The evaluation method of the classifier for feature evaluation usually is k-fold cross validation.

Recursive Feature Elimination

Recursive Feature Elimination is a wrapper that recursively reduces the set of features by eliminating the least important ones based on the ranking provided by a specific model.

- **Feature Dependency:** RFE takes into account the feature dependency assuming that are incorporated in the ranks.
- **Feature Ranking:** Not only does RFE help in selecting important features, but it also ranks all features based on their importance.
- **Handles Multicollinearity:** If some features are correlated, RFE can help in identifying and retaining the most important one among them.
- **Flexibility:** RFE can be used with any model that assigns weights or importance to features, making it versatile.

- **Computationally Intensive:** As RFE involves training the model multiple times (once for each feature), it can be computationally expensive.
- **Model Dependency:** The effectiveness of RFE is tied to the chosen model. A poorly chosen model might lead to sub-optimal feature selection.
- **Stability Issues:** Slight changes in the data can lead to different rankings, especially when features have similar importance.
- **Base Model Selection:** The choice of the model used in RFE is crucial. It's advisable to use a model that naturally provides importance or coefficient values, like decision trees, linear regression, or support vector machines.

Embedded Methods

White box predictors are actually based on feature selection. So, embedded methods are all the learning algorithms that derive white box predictors. These include: Decision trees and rules, and Shrinkage methods such as ridge regression and LASSO.

6 Model Validation

Lecture 7: Model Validation

Model Validation

Confusion Matrix

How to validate classifier performance? We use the following confusion matrix (table):

Actual/Predicted	**Pos**	**Neg**
Pos	TP	FP
Neg	FN	TN

Where:

- TP: True Positives
- FP: False Positives
- FN: False Negatives
- TN: True Negatives

Performance Metrics

Accuracy:

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

Error Rate:

$$\text{Error Rate} = 1 - \text{Accuracy} = \frac{FP + FN}{TP + TN + FP + FN}$$

Precision:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (Sensitivity, True Positive Rate):

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

Specificity (True Negative Rate):

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

F1-Score:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

ROC Curve

- ROC (Receiver Operating Characteristic) curve plots the True Positive Rate (Sensitivity) against the False Positive Rate (1 - Specificity) at various threshold settings.
- The area under the ROC curve (AUC-ROC) provides an aggregate measure of performance across all possible classification thresholds.
- A model with perfect discrimination has an AUC of 1.0, while a model with no discrimination has an AUC of 0.5 (a diagonal line).

Cross-Validation

K-Fold Cross-Validation:

1. Randomly split the dataset into k equal-sized folds (subsets).
2. For each fold k:
   - Use fold k as the validation set.
   - Use the remaining k-1 folds as the training set.
   - Train the model on the training set and evaluate it on the validation set.
3. Calculate the average performance across all k folds.

Stratified K-Fold: A variation that preserves the percentage of samples for each class in each fold.

Bias-Variance Tradeoff

- **Bias:** Error due to overly simplistic assumptions in the learning algorithm. High bias can cause underfitting.
- **Variance:** Error due to too much complexity in the learning algorithm. High variance can cause overfitting.
- The goal is to find the right balance between bias and variance to minimize total error.

Regularization

Regularization techniques help prevent overfitting by adding a penalty term to the loss function:

L1 Regularization (Lasso):

$$L(\mathbf{w}) = \text{Loss}(\mathbf{w}) + \lambda \sum_{i=1}^{n} |w_i|$$

L2 Regularization (Ridge):

$$L(\mathbf{w}) = \text{Loss}(\mathbf{w}) + \lambda \sum_{i=1}^{n} w_i^2$$

Hyperparameter Tuning

Grid Search: Exhaustive search over a specified parameter grid.
Random Search: Randomly samples parameter combinations from a distribution.
Cross-Validation: Used to evaluate each hyperparameter combination's performance.

Model Comparison

- Compare models using appropriate metrics (e.g., accuracy, F1-score, AUC-ROC).
- Use statistical tests (e.g., paired t-test) to determine if differences in performance are statistically significant.
- Consider computational efficiency, interpretability, and other practical aspects.