

Tutorial: The Elements of Statistical Learning II - SS06

1. Summary: Learning methods

For the learning methods and algorithms listed below, describe the purpose, possible applications and main advantages and disadvantages. When it is appropriate, judge the methods with respect to handling of missing values, robustness, runtime, interpretability and predictive power.

1. Local Polynomial Regression

Prediction method; in local neighborhoods polynomial model assumptions; tradeoff between linear regression and nearest-neighbors (NN); weaker assumptions than in linear regression, smoother solutions than with NN; bandwidth for kernel is sensitive parameter to be chosen or fitted; slower than linear regression; good predictive power if model assumptions hold.

2. Neural Networks

Prediction method; two-stage model, a nonlinear activation function is applied to linear combinations of the inputs to obtain hidden variables, the same type of transformations are applied to model outputs from the hidden variables; requires experience or prior data knowledge due to large number of parameters; even more flexibility through adaption of network topology; possible overfitting; low interpretability of black box process; difficulties with missing data; possibly high predictive power.

3. Support Vector Machines

Classification method; linear decision boundary is selected by maximizing margin between classes; misclassifications are penalized; transformation to higher dimensional space through kernel allows modelling of non-linearities; convex optimization problem computationally easy; due to kernel trick no knowledge of explicit data transformation required; choice of kernel crucial; especially useful for high-dimensional data; high predictive power.

4. Flexible Discriminant Analysis

Classification method; reformulates classification as regression problem; regression methods can be then be used for classification.

5. Mixture Discriminant Analysis

Classification method; each class is modelled as a mixture of Gaussian; more flexible generalization of LDA; number of Gaussians is adaptive parameter that can be difficult to optimize; assumes linear decision boundaries like LDA, thus rather strong model assumptions.

6. Discriminant-Adaptive Nearest Neighbor Method

Classification method; Generalization of Nearest Neighbor-Classification; selects most discriminating direction in neighborhood of target point and then adapts distance metric according to this direction before using Nearest Neighbor-Classification.

7. The Apriori Algorithm
Association rule mining method; searches for associated items in large databases; computationally fast; especially useful for huge databases; finds only rules with high support; lower support bound increases computation time drastically.
8. K-means Clustering
Unsupervised learning to find clusters in data; after finding initial cluster midpoints two steps are iterated until convergence: cluster assignments are obtained by minimizing Euclidean distances to the midpoints and midpoints are calculated as cluster means based on the assignments; converges to local optimum of objective function which is the sum of within-cluster variances; yields good results for similar cluster sizes and without outliers.
9. Vector Quantization
Dimension reduction and classification method; VQ represents building blocks by prototypes; Learning vector quantization (LVQ) iteratively selects training points and moves them towards one of a number of given prototypes.
10. K-medoids Clustering
Unsupervised learning to find clusters in data; like k-means with Euclidean distance replaced by different similarity measure; the minimization problem to find new cluster midpoints that minimize the objective function can be by far more computationally expensive, but more robust to outliers.
11. Hierarchical Clustering
Unsupervised learning to find clusters in data; agglomerative methods recursively merge at each level the pair of clusters with the smallest intergroup dissimilarity into a single cluster, divisive methods recursively split a cluster into two new clusters with a largest possible between-group dissimilarity; computationally fast and easy; provides dendrograms, which are intuitive, simplifying, summarizing plots.
12. Self-Organizing Maps
Unsupervised learning; like k-means but with additional topological constraints for the midpoints, which are modelled to lie on a one- or two-dimensional manifold.
13. Principal Component Analysis
Dimension reduction method; data are projected onto lower-dimensional space based on maximal variances across orthogonal directions; provides a sequence of best linear approximations to the data for any dimension smaller or equal than the original one; useful for preprocessing and visualization of data.
14. Principal Curves and Surfaces
Dimension reduction; generalizes PCA to nonlinear surfaces; iteratively fit a function that is average of all data points projected onto it.

2. Summary: Applications in bioinformatics

For the following algorithms or terms, briefly explain their principles and goals and describe which statistical learning methods they use.

1. PxPAM and PxKmeans

Cluster pixels of a microarray spot into two regions that represent foreground and background area; methods: PAM (Partitioning around medoids), K-means clustering.

2. Variance-stabilizing normalization

Normalization method for microarray data; render variance of expression values constant for different intensity values; method: Maximum-Likelihood for a linear regression model with additive and multiplicative noise term.

3. Gap statistic

Estimate optimal number of clusters in a gene expression data set; for different numbers of clusters compare the objective function with an average value obtained by randomly permuting cluster labels, select the number with the largest gap; works only well for low-dimensional data; statistical learning method: permutation test.

4. Average silhouette width

Estimate optimal number of clusters from a distance matrix; calculate silhouette values of observations by relative distance to their cluster prototype and the closest of the other cluster prototypes, minimize the sum of these values to select the number of clusters.

5. PAM (Prediction Analysis for Microarrays)

Classification method for microarray samples with known labels; method: regularized version of diagonal linear discriminant analysis, cross-validation is used to select regularization parameters.

6. ScorePAGE

Score gene sets defined by metabolic pathways for co-regulation in a gene expression data set; calculate average co-regulation in gene set and compare with scores of randomly selected sets; method: nonparametric permutation test.

7. ProtFun

Predict probabilities of membership of a protein to functional categories from features that are calculated from the amino acid sequence of the protein; iteratively, neural networks of increasing size are fitted and the best models based on test set performance are selected and combined to an ensemble network; method: ensemble of neural networks.

8. FunStruc

Predict structural class of a protein from the amino acid sequence of the protein; use ProtFun values as input data; for a fixed structural class: for every functional category in ProtFun, calculate maximum-likelihood intervals of values of class members and corresponding likelihood scores; for a query sequence, compare ProtFun values with

intervals and combine most significant hits to an ensemble score; method: weighted voting algorithm based on maximum-likelihood approach.

9. STRuster

Automated clustering of sets of proteins structures; cluster models according to structure similarity; compute Manhattan distance between corresponding $C\alpha$ distanced; use filters for distances for regularization; methods: hierarchical clustering for visualization, PAM (Partitioning around medoids) for clustering, average silhouette width for estimating optimal number of clusters.

10. mtreemix

Modelling of complex multivariate data; use mixture of mutagenetic trees to approximate multivariate distribution of binary events; model selection via cross-validated log-likelihood; method: maximum weight branching algorithm for single tree; EM-algorithm for the estimation of mixture of trees; bootstrap analysis for estimating tree stability.