Вычислительная статистика

1 Краткое описание курса

2 Предварительные требования

Для успешного освоения материала курса требуется владеть основами программирования, математического анализа, линейной алгебры и геометрии, математической статистики. Знакомство с языком программирования Python желательно, но необязательно.

3 Получаемые компетенции

По итогам освоения курса студент должен:

1. уметь выбирать ПО для анлиза данных
2. знать методы оптимизации

4 Содержание курса

(pg. 147–567)

1. **Введение** (2)
   Библиотека машинного обучения scikit-learn.
2. **Supervised learning** (4)
   3.1.1 Generalized Linear Models Ordinary Least Squares Ridge Regression Lasso Using cross-validation Information-criteria based model selection Multi-task Lasso Elastic Net Multi-task Elastic Net Least Angle Regression (LARS) LARS Lasso Orthogonal Matching Pursuit (OMP) Bayesian Regression Bayesian Ridge Regression Logistic regression Stochastic Gradient Descent - SGD Perceptron Passive Aggressive Algorithms Robustness regression: outliers and modeling errors RANSAC: RANdom SAmple Consensus Theil-Sen estimator: generalized-median-based estimator Huber Regression Polynomial regression: extending linear models with basis functions

   3.1.2 Linear and Quadratic Discriminant Analysis Dimensionality reduction using Linear Discriminant Analysis 3.1.3 Kernel ridge regression 3.1.4 Support Vector Machines Classification Multi-class classification Scores and probabilities Unbalanced problems Regression Density estimation, novelty detection Complexity Kernel functions

   3.1.5 Stochastic Gradient Descent Classification Regression Stochastic Gradient Descent for sparse data Complexity

3.1.6 Nearest Neighbors Unsupervised Nearest Neighbors Finding the Nearest Neighbors KDTree and BallTree Classes Nearest Neighbors Classification Nearest Neighbors Regression Nearest Centroid Classifier Nearest Shrunken Centroid

3.1.7 Gaussian Processes Gaussian Process Regression (GPR) GPR with noise-level estimation Comparison of GPR and Kernel Ridge Regression Gaussian Process Classification (GPC) Probabilistic predictions with GPC Kernels for Gaussian Processes Radial-basis function (RBF) kernel Matérn kernel Rational quadratic kernel Exp-Sine-Squared kernel Dot-Product kernel

3.1.8 Cross decomposition

3.1.9 Naive Bayes Gaussian Naive Bayes Multinomial Naive Bayes Bernoulli Naive Bayes Out-of-core naive Bayes model fitting

3.1.10 Decision Trees Classification Regression Multi-output problems Tree algorithms: ID3, C4.5, C5.0 and CART Classification criteria Regression criteria

3.1.11 Ensemble methods Bagging meta-estimator Forests of randomized trees Random Forests Parallelization Feature importance evaluation AdaBoost Gradient Tree Boosting

3.1.12 Multiclass and multilabel algorithms

3.1.13 Feature selection Removing features with low variance Univariate feature selection Recursive feature elimination Feature selection using SelectFromModel L1-based feature selection Tree-based feature selection

3.1.14 Semi-Supervised Label Propagation

3.1.15 Isotonic regression

3.1.16 Probability calibration

3.1.17 Neural network models (supervised) Multi-layer Perceptron Classification Regression Regularization

3. **Unsupervised learning** (4)

3.2.1 Gaussian mixture models Gaussian Mixture Pros and cons of class GaussianMixture Estimation algorithm Expectation-maximization Estimation algorithm: variational inference The Dirichlet Process

3.2.2 Manifold learning Isomap Locally Linear Embedding Modified Locally Linear Embedding Hessian Eigenmapping Spectral Embedding Local Tangent Space Alignment Multi-dimensional Scaling (MDS) t-distributed Stochastic Neighbor Embedding (t-SNE) Optimizing t-SNE Barnes-Hut t-SNE

3.2.3 Clustering K-means Mini Batch K-Means Affinity Propagation Mean Shift Spectral clustering Different label assignment strategies Hierarchical clustering Different linkage type: Ward, complete and average linkage Adding connectivity constraints Varying the metric DBSCAN Birch Clustering