Review of COMS 4721

Key ideas

Key ideas (1)

Major themes

- Statistical models and the plug-in principle
- Linearity and features
- Inductive bias
- Optimization

Algorithms/methods

- Greedy heuristic for decision trees
- Cross validation
- Model averaging, bagging
- OLS, MLE, ERM
- Perceptron
- Ridge regression, SVM
- Gradient descent, SGD, autodiff
- PCA, SVD, power method
- AdaBoost

Key ideas (2)

Other important ideas

- Structure of optimal predictors
- Testing
- Bias-variance decomposition
- Multiple perspectives on linear regression and classification
- Regularization, data augmentation
- Surrogate losses
- Convexity (at a rudimentary level)
- Objectives: costs, calibration, fairness
- Reductions
- Covariance matrices
- Low-rank matrix approximation
- Kernel trick
- Parameterized predictors / architectures
- Sampling bias, importance weights

Lecture-by-lecture review

Overview of ML

- Basic steps of supervised ML
- Cast of characters: feature vectors, labels, ML algorithm, predictor, etc.
- Features, predictor templates, training, testing

Decision trees

- Structure of decision trees
- Greedy heuristic
- Stopping criteria
- Occam's razor

Model selection

- Cross validation (hold-out method, K-fold cross validation)
- Domain-specific cross validation

Statistical models for prediction

- Coin toss model
- IID models for classification and regression
- Plug-in principle
- Structure of optimal classifiers
- Test error rate
- Bias-variance decomposition
- Approximation error and variability

Ensemble methoes

- Model averaging and its MSE
- Bagging, Bootstrap, Random Forests

Linear regression

- Ordinary least squares
- Normal equations
- Features expansion
- "Over-fitting"
- OLS in terms of orthogonal projections
- OLS as MLE for normal linear regression model
- OLS as ERM for linear predictors

Linear classification

- Logistic regression model
- Geometry of linear classifiers
- ERM for linear classifiers
- Perceptron

Inductive bias and regularization

- Minimum Euclidean norm solution to normal equations
- Ridge regularization
- Data augmentation
- Margins in linear classification
- Support vector machines
- Logarithmic losses / surrogate losses

Numerical optimization

- Best affine approximations and gradients
- Gradient descent
- Properties of gradient descent
- Convexity
- Stochastic gradient descent
- Autodiff

Classification objectives

- Handling costs in binary classification
- Probability calibration
- Fairness / COMPAS example

Reductions

- One-against-all reduction
- ECOC reduction

Multivariate Gaussians

- Covariance matrices and their basic uses
- Multivariate Gaussians, marginal distributions, conditional distributions

PCA / SVD

- Optimality properties of PCA
- Power method
- Low-rank matrix approximation and optimality of SVD
- Connection between SVD and PCA
- Latent Sematic Analysis
- Matrix completion

Kernel machines and neural nets

- Families of functions that yield "universal approximation"
- "Kernel trick" (i.e., the dot product trick for certain feature expansions)
- Implicit representations of linear functions in span of training feature vectors
- Kernel machines vs neural networks
- Neural networks as straight-line programs
- Practical issues for learning with neural networks

Boosting

- Boosting vs Bagging
- AdaBoost, basic properties
- Margins
- Weak classifiers in face decition
- Cascade architecture

Interactive learning

- Interactive learning vs non-interactive learning
- Sampling bias