Probability Theory

- probability measures, random variables
- expectation, variance
- conditional probabilities, conditional expectations
- discrete probability measures, probability densities
- Bayes formula

Decision Theory, Maximum Likelihood, and Discriminant Functions

- setting for supervised learning
- Bayes theorem
- decision rules, Bayes risk
- discriminant functions
- discriminant functions for special cases of Gaussian class densities
- Maximum Likelihood Estimation of parametric densities.

Principal Component Analysis

- PCA finds low-dimensional subspace which contains most of the variance of the data.
- Used for: dimensionality reduction, reducing the number of highly correlated variabels.
- Basic idea: Compute Eigendecomposition of the covariance matrix.
- Problems: How to decide on the number of embeddings ('knee'?)

Independent Component Analysis

- Problem: We observe a linear mixture of independent sources (for example, time-series data)

$$y(t) = Ax(t)$$

  where A is unknown. We want to compute the unmixing matrix A^{-1}.
- Applications: Blind Source Separation, removal of artifacts.
- TDSep: Joint diagonalization of covariance matrices at time point t and t and $t - \tau$.

K-Means Clustering

- Problem: Partition a set of points into *clusters* such that each cluster contains similar points.
- Algorith: EM-like iteration
  1. Find closest center for each point.
  2. Recompute mean given these assignments.
- Other variants are real EM-algorithm for mixtures of Gaussians.

Agglomerative Clustering

- Problem: Build a tree which represents a nested partitioning of a set of points.
- Algorithm: (Bottom-up) Cluster two points or clusters together according to some cost function
- Where to cut the tree?

Stability-Based Model Selection for Clustering

- Problem: Find the "correct" number of clusters.
- Algorithm:
  1. Run the algorithm several times with different starting values, or resample from the data.
  2. Compute differences between clustering solutions.
  3. Normalize by "random clusterer".
  4. Pick most stable number of clusters.
- Problems: Still no guarantee that clustering really makes sense.

Statistical Test Theory

- Statistical Tests, p-values, significance levels, null Hypothesis...

Supervised Learning, a bit of Learning Theory

- Formal setup with joint probability distribution on x and y.
- loss function, expected risk
- empirical risk, empirical risk minimization, consistency
- uniform convergence
- other error measures: FPR, TPR, ROC, Precision&Recall, etc.

Least Squares Regression

- Problem: Fit a hyperplane ($f(x) = x^\top w + b$) or a linear combination of basis functions ($f(x) = \sum_{j=1}^{d} w_j \psi_j(x)$) such that the squared error is minimal

$$\sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2$$

- Algorithm: Weight vector can be computed in closed form, for example $w = (X^\top X)^{-1} X^\top Y$.
- Variant: *Ridge Regression* regularizes by penalizing the norm of the weight vector.

Model Selection

- How to choose basis functions, regularization constant, etc.?
- *Cross Validation*: Iteratively split data, train on one part, test on the other, choose parameters which minimize the test error.
- **Reliable estimate of test error only on data which has not been used for training!**
- C_p*-statistic*: For some models, one can estimate the optimism of the training error.

Support Vector Machines

- Main idea: Learn hyperplanes which have a large margin. Statistically robust.
- Algorithm: Boils down to a quadratic opimization problem, usually solved in the *dual formulation*.
- *Kernel Trick*: Replace scalar products by a kernel function k. Amounts to mapping data into high-dimensional feature space non-linearly to increase expressive power of linear hyperplanes.

Kernel PCA

- Kernel PCA = PCA in feature space.
- Instead of the covariance matrix, kernel matrix is considered.
- Instead of principal components, the scalar product with the direction can be computed

$$f_i(x) = \langle \Phi(x), v_i \rangle.$$

Summary

- Background
  - Probability Theory
  - Statistial Testing
- Unsupervised Learning:
  - PCA
  - ICA
  - Clustering
  - Kernel PCA
- Supervised Learning:
  - A bit of learning theory
  - Least Squares Regression
  - Support Vector Machines