Supervised Learning Revision Notes (18-19)

Study Suggestions

- Lecture notes
- Problems in lectures notes
- Past exams
- Assumed background knowledge includes but is not limited to
  1. Probability (Bayes rule, conditional probability, expectation, random variables, basic combinatorics)
  2. Linear Algebra (singular value decomposition, positive semi-definite, positive definite, rank, linear systems of equations)
  3. Misc: convexity, boolean functions (and, or, not, conjunctive normal form, disjunctive normal form, conjunction, disjunction)

Exam Format

Ten questions each with two sub-parts (each sub-part is 5 points) (answer all questions).

There are nine lecture files on moodle. The lecture "7. Sparsity and Matrix Estimation" is not explicitly examined. Each of the remaining 8 lectures has a question associated with it. The remaining two questions are also drawn from the 8 examinable lectures.

Lectures

DISCLAIMER: Exam is not limited to outline topic headers.

1. Introduction
   - Supervised learning model
   - Least squares
   - Introducing a bias term
     - Normal equations
     - Bayes Estimator
   - k-NN
     - 1-NN is asymptotically 2 $\times$ "optimal"
     - k-NN is optimal
   - Optimal supervised learning
   - Bias-variance decomposition
   - NFL Theorem
   - Hypothesis space
   - Bayes classifier

- Overfitting and Underfitting
- Cross-validation

2. Kernels and Regularization
   - Inner product/vector/normed space
   - Ill-posed problems
   - Ridge regression (as an example of regularisation)
   - Primal vs Dual representation
     - Computational considerations
     - Representer theorem
   - Feature maps
     - Basis functions - explicit feature map
     - Kernel functions - implicit feature Map
       * Definition (Role of PSDness)
       * Kernel construction
       * Example kernels : Polynomial, Anova, Gaussian
       * min Kernel
3. Tree-based learning algorithms and Boosting
   - Classification and Regression Trees
     - Recursive Binary Partition
     - Optimization formulation
     - "Greedy" approximate algorithm
     - Cost-complexity pruning
     - Classification trees
     - Node impurity measures
   - Ensemble Methods (Wisdom of crowds)
     - **Not examined 18-19:** Details of Chernoff bound argument
   - Bagging
   - Random Forests
   - Weak Learners
     - Definition
   - Boosting (Adaboost)
     - Weak Learner
     - Distribution on training set
     - Final classifier is a linear combination of weak classifiers
     - Exponential convergence of training error
     - Boosting as exponential minimiser
     - Boosting generalisation guarantees [**not examined 18-19**]
     - Additive Models, Exponential Loss (vs other loss functions) and Boosting
   - Comparison between boosting and bagging
4. Support Vector Machines
   - Linear Classifier
   - Hyperplane (Separating)
   - Margin of hyperplane and a point
   - Optimal Separating Hyperplane (OSH) (parameterization normal vs canonical)
   - Solution form of OSH in primal and dual (Combination of support vectors)
   - Support vectors and generalisation
   - Non-separable case
   - Role of the parameter C
   - connection to regularisation
5. Online learning I

- Online learning model
  - Loss bound
- Learning with expert advice
  - Halving algorithm
  - Weighted majority algorithm
  - Regret bound
  - Experts algorithm (AKA Weighted average algorithm) bound for general loss functions difference in results log and arbitrary loss function
  - Expected loss bound for WAA/Hedge
- Learning with thresholded linear combinations
  - Linear classifiers and disjunctions
  - Perceptron
  - Winnow
  - Learning boolean functions
    * Definitions (conjunction, disjunction, (monotone) literal, term, etc)
    * Perceptron and Winnow mistake bounds
    * Case study: Finding a maximally sparse classifier is NP-hard **[not examined 18-19]**
    * Case study: DNF
    (a) Anova Kernel
- Learning with sequences of experts **[not examined 18-19]**
- Tracking the best expert **[not examined 18-19]**
  - Fixed Share algorithm **[not examined 18-19]**
  - Shifting loss bound **[not examined 18-19]**

6. Sparsity and Matrix estimation **[not examined 18-19]**
7. Learning Theory
   - learning model
   - definitions of expected (AKA true error, generalisation error) and empirical errors
   - validation set bound
   - empirical risk minimisation (ERM)
   - "expected" vs "confident" bounds
   - PAC Model
     - Realisability assumption
     - role of ϵ and δ
     - NFL lower bound result
     - Learning with finite hypothesis classes
     - Sample complexity
   - VC-dimension (Definition as well as be able to compute for a hypothesis class)
   - VC-dimension (Large Margin Halfspaces)
   - VC-dimension upper bound for PAC learning and connection to finite hypothesis class
   - Agnostic model
   - Error decomposition approximation and estimation error.
8. Advanced Online Learning
   - Partial feedback setting
   - Motivation "exploration vs exploitation"
   - Unbiased estimator
   - Importance weighting
   - EXP3
     - Connection to hedge
     - Model : Deterministic Oblivious Adversary

- Theorem (bound how does it compare to hedge)

- Matrix completion
- Factor Model
- Rank Complexity, Margin Complexity
- mistake bound for matrix winnow applied to matrix completion
- Multi-task interpretation
- (k, ℓ)-biclustering (definition, VC-dimension lower bound, connection to margin complexity)

9. Graph-based Semi-supervised learning

- Overview
  - Why SSL?
  - Comparison to SL and UL
  - Transduction and Induction
- Graphs
  - Intrinsic vs extrinsic
  - How to build (k-NN, ϵ-ball, tree-based, weighted graph, combo)
  - Graph classifier
    * Cut as a measure of smoothness/complexity
- Algorithmic frameworks
  - Minimum cut
  - Laplacian
  - Spectral clustering (cut versus ratio objectives)
  - Interpolation as a limit case of regularization
- Minimum cut transduction
- Laplacian-based transduction
  - quadratic form $\mathbf{u}^T L \mathbf{u}$ (connection to cut)
  - associated kernel as pseudo-inverse
- Laplacian Interpolation (AKA harmonic minimization, label propagation, Laplacian interpolated regularization)
  - Motivation via consensus
  - Harmonic solution
- Interpreting Laplacian-based transduction
  - Graph as a resistive network
  - Effective resistance
    * Computation
    * Kirchoff Circuit Laws **[not examined 18-19]**
    * Connection to kernel (pseudo-inverse of Laplacian)
    * Proof that $R(i,j) := (\boldsymbol{e}_i - \boldsymbol{e}_j)^T L^+ (\boldsymbol{e}_i - \boldsymbol{e}_j)$ **[not examined 18-19]**
    * Connection to random walks
    * Labeling respects cluster structure (two-clique example)
- Sections VIII-X **[not examined 18-19]**