Supervised Learning Revision Notes (21-22)

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. Calculus (Integration and differentiation with multiple variables)
    4. Misc: convexity, boolean functions (and, or, not, conjunctive normal form, disjunctive normal form, conjunction, disjunction)

Exam Format

Eight questions (answer all questions).

There are eight lecture slide sets on moodle. Each question is associated with a lecture.

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
    - Curse of dimensionality
    - Hypothesis space
    - Bayes classifier

- Overfitting and Underfitting
- Cross-validation

2. Kernels and Regularization
   - Inner product/vector/normed space
   - Convexity
   - 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
     - Regularisation-based learning algorithms
       - Definition (Role of PSDness)
       - Kernel construction
       - Example kernels : Polynomial, Anova, Gaussian
       - min Kernel
   - Regularisation-based learning algorithms
3. Support Vector Machines
   - Linear Classifier
   - Hyperplane (Separating)
     - Paramerisation (normalized, canonical)
   - Margin of hyperplane and a point
   - Constrained optimisation with a Lagrangian
   - 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
4. Decision Trees and Ensemble Learning
   - 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)
   - 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 21-22**]

    - Additive Models, Exponential Loss (vs other loss functions) and Boosting
  - Comparison between boosting and bagging

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
    - Weighted Average Algorithm - Proof
    - Expected loss bound for WAA/Hedge
    - Hedge Theorem - Proof
  - Learning with thresholded linear combinations
    - Linear classifiers and disjunctions
    - Perceptron
      * Perceptron Bound and Proof [Novikoff]
    - Regret bounds for linear separation
      * Regularisation for batch vs online
      * Hinge loss
      * OGD Algorithm and derivation
      * Regret Bound (and proof) of OGD
      * Connections to perceptron
    - Winnow
    - Simple conversion of online to batch algorithm and bound
    - 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 21-22]**
      * Case study: DNF
      (a) Anova Kernel
  - Learning with sequences of experts (This and below discussed in "Lecture 8")
  - Tracking the best expert (Lecture 8)
    - Fixed Share algorithm (Lecture 8)
    - Shifting loss bound (Lecture 8)
      * Proof Sketch (Lecture 8)

6. 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
  - Graph Laplacian
    - quadratic form $\mathbf{u}^T L\mathbf{u} = \sum_{(i,j)\in E(G)} w_{ij}(u_i - u_j)^2$ (connection to cut)
    - In a connected graph $\mathbf{1}$ is the only eigenvector with eigenvalue 0.
    - How to build (k-NN, ϵ-ball, tree-based, weighted graph, combo)

- Spectral Clustering
  - Cut and RatioCut
  - RatioCut is a "hard" discrete optimisation problem
  - How the second eigenvector provides a "relaxed" approximate solution
- 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 20-21**]
    - Connection to kernel (pseudo-inverse of Laplacian)
    - Proof that $R(i,j) := (\boldsymbol{e}_i - \boldsymbol{e}_j)^T L^+ (\boldsymbol{e}_i - \boldsymbol{e}_j)$
    - Connection to random walks
    - Labeling respects cluster structure (two-clique example)
- Sections VIII-X [**not examined 20-21**]

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 (proof not examined)
     - 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 lower/upper bound for PAC learning and connection to finite hypothesis class (proof intuitions not examined)
   - Agnostic model
   - Error decomposition approximation and estimation error.
8. Online Learning II
   - Partial feedback setting
   - Motivation "exploration vs exploitation"
   - Unbiased estimator
   - Importance weighting
   - EXP3
     - Connection to hedge
     - Model : Deterministic Oblivious Adversary
     - Theorem & Proof (how does it compare to hedge)
   - Model : Online Multitask Learning with Long-Term Memory
   - Learning with sequences of experts (ie, the single task special case with experts)
   - Tracking the best expert
   - Fixed Share algorithm
   - Shifting loss bound
     - Proof Sketch
   - p43 – to end of slides ("Introducing memory") [**not examined 21-22**]