DS 4350, CS 5350/6350: Machine Learining Spring 2024

Final Review

April 18, 2024

What have we seen

This document summarizes the topics we have seen in the second half of the semester.

Learning theory

- Assumption that train and test examples are drawn from the same distribution
- How the batch model of learning different from mistake bound learning?
- PAC learning: How close will the approximation of the concept be (ϵ)? How sure are we that the learning algorithm will find a hypothesis with a good ϵ (δ)?
- Definition of PAC learning, efficient PAC learning
- Sample complexity and computational complexity
- Occam's razor for consistent learners (requires us to count sizes of hypothesis classes)
- Positive and negative results from sample complexity perspective. From computational complexity perspective.
- Agnostic learning. What is it? Using Hoeffding's bound to prove sample complexity result
- Infinite hypothesis spaces, shattering and VC dimensions
- Proving the VC dimension for simple function classes
- How VC dimension operates like size of the hypothesis space with respect to sample complexity, but for infinite hypothesis spaces.

Boosting and Ensembles

- What is Boosting?
- What is are weak and strong PAC algorithms?
- AdaBoost. The theoretical question it answers via a constructive proof, the algorithm itself, how it works, the final hypothesis it produces, training and test errors for Adaboost.
- What is an ensemble method? Bagging and random forests

Support Vector Machines

- Margins and VC dimensions, why large margins imply better generalization.
- Maximizing margins, geometric and functional margins, the geometric interpretation of maximizing margins, why margin maximization is equivalent to minimizing the norm of the weight vector
- Hard and soft SVMs, slack variables, the SVM objective, the hinge loss
- Optimizing the SVM objective, how gradient descent works, how stochastic sub-gradient descent works, why it is a better alternative than gradient descent, convergence and learning rates

Loss minimization

- The principle behind empirical loss minimization
- Why regularization is important
- The 0-1 loss
- Different surrogate functions for the 0-1 loss: Hinge, perceptron, logistic
- Interpreting the SVM objective as regularized loss minimization
- Hinge loss vs. Perceptron loss. How perceptron is just SGD with Perceptron loss.
- How regularization has a probabilistic interpretation as well (MAP estimation for logistic regression)

Bayesian Learning

- Using a probabilistic criterion to pick a classifier (vs. learning a probabilistic classifier)
- Bayes rule and how it applies to hypotheses and data, priors, likelihoods and posteriors
- Maximum A Posteriori learning
- Maximum Likelihood learning and how it relates to MAP learning
- Maximum likelihood estimation for simple probabilistic models. Maximizing log-likelihoods, using the i.i.d assumption of training examples to simplify $P(\text{data}|\text{hypothesis})$ as a product over the examples.
- How least squares regression is an instance of MLE on a specific probabilistic model, deriving least squares regression from the probabilistic perspective

Logistic Regression

- Predicting probabilities to label assignments instead of labels
- The sigmoid function, the logistic regression model
- Logistic regression and linear classifiers
- Using Bayesian criteria for training logistic regression: MLE for logistic regression, and MAP learning where weights are assumed to be drawn from a normal distribution. Deriving the objective in both cases
- How the MAP learning setting is equivalent to loss minimization, the logistic loss
- Deriving the gradient of the learning objective for logistic regression and instantiating stochastic gradient descent for this objective.

Neural Networks

- What is a multilayer neural network? What is an activation function? Popular activation functions.
- Expressiveness of two and three layer threshold networks
- How the forward pass works?
- The underlying concept behind backpropagation, how the backward pass fits in with SGD based learning.
- Training neural networks with SGD
- Practical concerns: Mini-batches, learning rates, dropout, how to pick hyperparameters

Discriminative and Generative models

- The definitions, how they differ, and why
- Examples of discriminative models
- Examples of generative models

Practical issues

- How a model is defined via features, hyper-parameters, loss and the actual parameters themselves.
- Choosing hyper-parameters via cross validation, how cross validation works, why it is important
- A high level understanding of the open questions involving fairness, accountability and transparency of machine learning systems.