Statistical Learning

Probability Distribution

* A specification of a probability for each event in our sample space

Joint Distribution

* Specification of probabilities for all combinations of events
* Marginalization

Conditional Probability

* Fraction of words in which B is true that also A is true.

Inference

Bayes Rule

* Pr(B|A) = Pr(A|B)Pr(B)/Pr(A)
* Prior: Pr(H)
* Likelihood: Pr(e|H)
* Evidence: e=<e1, e2, ….eN>

Bayesian Learning

* Amounts to computing the posterior using Bayes’ Theorem:

Pr(H|e)= kPr(e|H)Pr(H)

* No overfitting
* If the hypothesis space is large, Bayesian learning is intractable.
* Sum over hypothesis is often intractable. Thus, we need to approximate.

Bayesian Prediction

* Predictions are weighted averages of the prediction of individual hypotheses.
* Hypothesis serves as “intermediaries” between raw data and prediction.

Maximum a posterior

* Make a prediction based on the most probable hypothesis hmap.
* Less accurate than Bayesian since it only replies to one hypothesis.
* But map and Bayesian coverage as data increase.
* Controlled overfitting
* Finding hmap may be intractable.

Maximum Likelihood

* Simplify map by assuming uniform prior.
* Make prediction based on hML
* Less accurate than Map and Bayesian since it ignores the prior info and reply only on one hypothesis.
* But three coverages as data increase
* Subject to overfitting
* Finding hML is easier than hmap.

Lecture 10: Linear Regression MLE

Noisy Linear regression

* Gaussian Noise
* Find best w by max the likelihood of the data.

General Formulation (OLS)

* Cost function
* Objective
* OLS and MLE lead to the same optimal set of coefficients.
* The MAP is similar to MLS with L2 regularization.

Generalized linear model

* Classes are not separable with the feature
* We can do a nonlinear classifier or add features.

Lecture 11: Logistic Regression

Logistic regression

* Models the relationship between one or more indep variables and one binary dependent variable.
* the independent variables have a linear relationship with the logit transformation of the dependent variable
* Logistic function
* The coefficient tells you what change to expect in the *log odds ratio* of your dependent variable, for a one-unit increase in your independent variable

Support Vector Machine

* A separate plane represented by a weight vector w and a intercept b
* Overfitting
* Max the margin

Soft-Margin SVM

* No separate plane
* Introduce penalties to mis-classifications
* Helps prevents overfitting

Lecture 12: Naïve Bayes

Spam Classification

* about what characteristics distinguish spam from non-spam
* each word in the email as a feature
* We are trying to learn a function f(x) that maps input data xi, consisting of terms t1,…,tk, to predicted class
* X: x1…xn: n emails
* Xi: t1….tk: each email xi is comprised of k features of tj.

Naïve Bayes (Indep)

Maximum A-Posteriori (MAP) estimate

* Easier to work with log
  + Log function is monotonic – whatever maximizes x maximizes log(x)
  + Logs avoid floating point issues

Spam Classification Parameters

* A simple distribution over words is also called a unigram distribution
* commercial applications
* parameters are so easy to estimate.
* But it does not do the normalization.
* Relative likelihood matter

Smoothing

* avoid zero probability estimates

Laplace Smoothing

* Pretend we saw every outcome 1 more time than we actually did

Additive Smoothing

* Cross-validation

Lecture 13: Decision Tree

Forward Selection

* Add features to model

Decision Tree

* Build a tree from training data to explain label example
* Progressively add features that are most informative
* Keep the tree as small and as simple as possible
* decision trees are limited to these rectangular partitions

Hypotheses Space

Issue

* smallest tree
* reduce overfitting

Overfitting

* Stop growing tree when split is not statistically significant
* Grow tree, then prune afterwards
* Set maximum depth

Pruning

* Starting with the deepest nodes, delete splits where the value of the split does not exceed some threshold, T

Multiple-valued features

* turn a single feature into many binary features.

Continuous Features

* Bucket or threshold values

Regression Tree

* Output values are continuous.

Lecture 14: Random Forest

Random Forest

* Combine outputs of multiple classifiers
* No pruning
* Regularization

Ensemble Methods

* Bagging = bootstrap aggregating
* Stacking = train models on the output of other models
* Boosting = Train a sequence of models, each emphasizing the examples misclassified by the previous model
  + Individuals often make mistakes, but the “majority” is less likely to make mistakes.

Boosting

* Can boost any weak learning algorithm

Adaptive Boosting

Gradient Boosting

* Fits on residuls