Course outcomes

What you should know

- Review of all the previous lectures with expected outcomes from each lecture

Lecture 1

- MLE and MAP for coin flips: estimate the probability Θ of a coin landing on heads
  - MLE:
    - Notion of likelihood function
    - Setting up the objective function
    - Verifying the function is concave
    - Maximizing the objective function

Lecture 1

- MAP:
  - Notion of prior and posterior functions, and their relationship with the likelihood through Bayes rule.
  - Using the Beta prior
  - Setting up the objective function
  - Verifying the function is concave
  - Maximizing the objective function
  - How does the prior affect the final outcome? How can it be interpreted in terms of additional throws?

Lecture 2

- Familiarity with the notion of conjugate priors:
  - In the coin flip MAP example, the beta distribution is the *conjugate prior* of the Bernoulli likelihood function. We also call the prior and the posterior *conjugate priors*
- Probability review:
  - Probability distribution functions (PDF) for continuous variables
  - Expectation, Variance
  - Joint distributions, chain rule, Bayes rule.

Lecture 3

- Decision rules
  - Bayes decision rule
    - Bayes error/unavoidable error
- Definition of generative and discriminative classifiers

- KNN:
  - How to use KNN?
  - Training error vs test error.
  - How to pick K by cross-validation (what is cross-validation)
  - Behavior of the algorithm when K is small and when K is large, in terms of model complexity and risk of overfitting.
  - Behavior of the algorithm when K=1
  - What does overfitting mean?

Lecture 4

- Linear regression
  - Mean squared error statement and derivation
  - Probabilistic statement and derivation, and equivalence of the two solutions (from homework)
- Ridge regression
  - Mean squared error and ridge penalty statement and derivation
  - Probabilistic statement and derivation, and equivalence of the two solutions (from homework)
- Lasso regression:
  - What does it enforce, how do the learned parameters change with increased penalty?

Lecture 4

- Bias variance tradeoff:
  - The expected loss when the model is correctly specified can be decomposed into the sum of three components: bias^2, variance and noise (no need to memorize the derivation from class)
  - You need to know how to compute the bias of an estimator, and what it corresponds to
  - You need to know how to compute the variance of an estimator, and what it corresponds to

Lecture 5 - Naive Bayes

- Conditional independence assumption of naive bayes. How does it it help with learning less parameters?
- Naive Bayes Algorithm with binary X and Y:
  - How is it specified?
  - How to train it?
  - How many parameters does it require for the user to learn?
  - How to use with for prediction?
  - What is chance performance?
  - What happens when the Xs are not conditionally independent? What happens when some of the variables are irrelevant?

Lecture 5 - Naive Bayes

- Naive bayes for text classification:
  - What is the bag of word model?
    - How to formulate it?
    - How to learn it?
  - How to deal with words what we never encounter? What kind of prior can we use?
- Gaussian naive bayes (GNB):
  - What are the assumptions behind GNB?
  - How does it account for continuous variables?

Lecture 6 - Logistic Regression

- Logistic regression:
  - Logistic regression outputs the probability of Y given X.
  - How is the problem setup?
  - It doesn't have an analytical solution so gradient descent/gradient descent has to be used to learn the weights, depending on the formulation of the problem
  - What is block gradient descent? What is stochastic gradient descent? What are their characteristics?
  - What is the learning rate? What are the drawbacks/ benefits of a large learning rate or a small learning rate?

Lecture 6 - Logistic Regression

- Logistic regression:
  - Without regularization, the logistic regression problem is ill specified and the magnitude of the weights will tend to infinity as the probability of the training labels is maximized. Regularization offers a tradeoff between weight size and training error
  - How many parameters need to be estimated?
- Logistic regression (LR) vs GNB:
  - If the variances learned for each X_i are made to be equal across classes, GNB learns a linear decision boundary.
  - GNB converges faster to its asymptotic error
  - If the variances are equal across classes and if the X_is are conditionally independent given Y then GNB and LR perform similarly.
  - GNB might perform worse than LR if the data is not conditionally independent but it is not always easy to predict

Lecture 7 - Decision trees

- How to use a decision tree to make predictions
- Entropy, conditional entropy and information gain statements.
- How to use information gain to greedily build a tree (using ID3). This approach optimizes the length of the tree to obtain short trees.
- How to interpret train error / test error plots. What does overfitting correspond to mathematically/graphically?

Lecture 8 - Perceptron and NN

- The perceptron algorithm:
  - Statement, how to train it, how to use it to predict
  - The effect of a misclassified positive sample in training: changes the training prediction in the positive direction.
  - The effect of a misclassified negative sample in training: changes the training prediction in the negative direction.
  - The notion of margin and of a linearly separable training set.

Lecture 9 - Neural Networks

- Neural networks:
  - Stacking layers with non-linear activations allows us to learn complex decision boundaries
  - Sigmoid functions are used instead of sign functions because they are differentiable. They however introduce problems such as the vanishing gradient problem.
  - The steps required to train a neural network: (We saw this in detail in HW and in class)
    - Including how to use back-propagation to estimate the gradient at every parameter of the network.

Lecture 9 - Neural Networks

- Different tricks can be used to optimize neural networks, leading to a large array of decision involving the number of layers, the number of nodes, the choice of different activation gates or of different loss functions.
- Neural networks can be used in order to learn transformations of input data such as images and sounds into a space in which they are characterized by features that are important for the task at hand. This is referred to as representation learning.

Lecture 10/11 - SVMs

- Separable training set - linear SVMs:
  - Problem statement
  - Writing down the Lagrangian then the dual formulation and the solution of the dual problem.
  - How to use a trained SVM to make predictions. What decision boundaries are learned?
  - The dual formulation lets us see that the algorithm relies on the dot product of new test points with the support vectors.
- How to show that k(x,y) is a kernel?
  - Mercer theorem or expressing k(x,y) as a dot product of the same feature map of x and y.

Lecture 10/11 - SVMs

- Kernel SVMs:
  - We can use the kernel trick to allow us to efficiently train and use SVMs with non-linear kernels, allowing us to learn complex decision boundaries
  - Common kernels and resulting decision boundaries (from homework)
- Non-separable training set:
  - We can use SVMs with slack variables to learn decision boundaries (homework)

Lecture 12 - Boosting

- Boosting uses weak learners with accuracy above 0.5 percent to derive a strong learner
- Adaboost algorithm:
  - How does it works for prediction / what does it output?
  - How is it trained?
  - It ends up performing surprisingly well in some scenarios. It could suffer from overfitting sometimes, and sometimes the test error keeps improving even when the training error is zero.

Lecture 13/14 Learning Theory

- Notion of the complexity of the hypothesis space H.
- Theory to relate the number of training examples, the complexity of hypothesis space, training error and true error (as well as how training examples are presented).
- Mathematical formulation of overfitting
- The notion of ε-exhausting a version-space
- The notion of VC dimensions and how to derive it with simple classifiers in the case where H is not finite

Lecture 13/14 Learning Theory

With probability $\geq (1-\delta)$, $(error_{true} - error_{train}) \leq \epsilon$

(1) for all $h \in H$ such that $error_{train} = 0$,

$$\epsilon = \frac{\ln|H| + \ln(1/\delta)}{m}$$

finite H

(2) for all $h \in H$

Agnostic

$$\epsilon = \sqrt{\frac{\ln|H| + \ln(1/\delta)}{2m}}$$

finite H

(3) for all $h \in H$

Agnostic

$$\epsilon = 8\sqrt{\frac{VC(H)(\ln\frac{m}{VC(H)} + 1) + \ln(8/\delta)}{2m}}$$

infinite H

Lecture 13/14 Learning Theory

- In the case where we have consistent classifiers, the number of examples we need grow as a function of $\frac{1}{\epsilon}$
- In the case of agnostic learning (training error is not 0) the number of examples we need grow as a function of $\frac{1}{\epsilon^2}$

Lecture 17 - k-Means

- What is clustering in general
- What is the objective of k-Means? What results does such an objective make likely?
- Understanding Lloyd's Method.
- What is the effect of the initialization on the final results? What are possible failure modes? What are solutions?

Lecture 18 - EM and GMMs

- Probabilistic Expression of a GMM
- Difference between k-Means and GMM solutions
- Using the EM algorithm (how it works in general)
- Adapting the EM algorithm to GMMs

Lecture 19-20 - Graphical Models

- Review of conditional independence and marginal independence
- Definition of Bayes Nets and how they represent probability distributions. How to derive the joint distribution from looking at a Bayes net. How to read conditional independence relations from a graph.
- Difference approaches for using a Bayes Net for inference and for learning a Bayes Net.

Lecture 21 - HMMs

- Understanding the setup of HMMs, what is known (observations) and what underlying latent variable model is assumed
- Being able to state the Markov assumption (for state transitions) and independence assumptions (for observations conditional on state).
- Using dynamic programming to calculate the probability of observations given the state transition probability matrix and observation probability matrix.

Lecture 22- SVD

- The definition of the singular value decomposition: the properties of the 3 matrices in the SVD, eg: singular values are always positive.
- The interpretation of SVD in terms of providing the best rank k approximation to a matrix in Frobenius norm.
- Interpreting the singular values, the meaning of an “approximately low rank” matrix.
- Understand why the SVD is not unique when two singular values are equal (for example the identity matrix does not have a unique SVD)

Lecture 23 - PCA

- Understand the different purposes of dimensionality reduction: computational, statistical, interpretability, and storage.
- Deriving PCA in terms of the direction that maximizes variance of the projected data.
- Deriving PCA in terms of the direction that minimizes reconstruction error.
- Writing PCA as finding the first eigenvalue and eigenvector of the covariance matrix.
- Understand the relationship between SVD and the eigenvalue decomposition for square symmetric matrices.

Lecture 24-25 - Reinforcement Learning

- Understand the problem setup of an infinite horizon MDP, in terms of states, rewards, actions, value, discount and so on.
- Be able to state Bellman's equation(s) for the value of the MDP and justify what it means.
- Use Bellman's equation(s) to solve an LP to get the value of each state.
- Derive the value iteration algorithm
- Derive the policy iteration algorithm

Lecture 26-27 - Sources of bias in applied ML

- Understanding the difference between internal and external validity, and different sources of bias and variance in data analysis.
- Comparing classifiers, estimating variance, using error bars.
- Understanding how multiple hypothesis testing or multiple comparisons inflates the error rate, and correcting for it.
- Sampling bias, and how it applies to predicting election outcomes

Lecture 26-27 - Sources of bias in applied ML

- The Kaggle competition "paradox" and overfitting to the validation set.
- What happens when the noise is not independent of the covariates, or of the ground truth.
- Understanding Simpson's paradox, and the UC Berkeley admissions example.
- Other sources of bias: heteroskedastic noise, covariate shift, label shift, non-iid data, etc.