Alin Tomescu
6.867 Machine learning | Week 7, Thursday, October 17th, 2013| Lecture 12

Lecture 12

Exam at 7pm. Location by last name letter:

- Room 1-190 (A-J)
- Room 6-120 (K-P)
- Room 35-225 (Q-Z)

Type of things we should know about the topics.

Basics of prediction learning

We are trying to solve forecasting problems, important to understand difference between this and an optimization problem. Understanding doing well on the training set does not mean you'll do well on the test set. What does generalization error mean? Large pool of examples, training set drawn from there, but you are tested on the entire pool.

Linear classifiers

- Definition. With or without offset (what they can or cant do).
- Decision boundary with or without features
- Understand kernel interpretation
  - $\sum_i \alpha_i y_i K(x_i, x)$
  - $\hat{\theta} = \sum_i \alpha_i y_i \phi(x_i)$
- Margin is important, be able to calculate it, either in the primal form or in the dual
  - How far is this point from the decision boundary implied by the kernel in the feature space
    - Need to know what is $\|\phi(x)\|$ and $\left\|\hat{\theta}\right\|$
    - No clear notion of margin in the normal space, ONLY in the feature space implied by the kernel space
  - $\frac{1}{\|\hat{\theta}\|} \to 0$ then $boundary \to \infty$

Perceptron algorithm

- You have to know it. They may give it to us.
- Properties
  - May not always converge
  - What is the meaning of saying that there's a bound on the number of mistakes R^2/γ^2
    - What is R?
- Traverse between primal/dual form of these classifiers
  - With/without kernel

Maximum margin separators: SVM

- What is the value of the margin: $\frac{1}{\|\hat{\theta}\|}$

Alin Tomescu

6.867 Machine learning | Week 7, Thursday, October 17th, 2013| Lecture 12

- margin boundaries: what do they mean? How we formulated the MMLC in terms of these margins
  - For $\|\hat{\theta}\| = \sqrt{\theta^2} = \sqrt{(\sum_i \alpha_i y_i \phi(x_i))(\sum_i \alpha_i y_i \phi(x_i))} = \sqrt{\sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j)}$
- Understand with or without kernels
- Primal/dual for finding MMLC
- Given dual formulation: $\sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j)$, s.t. $\alpha_i \geq 0$
  - Expected to know what does $\alpha_i \leq c$ would do?
  - Or if we add constraint $\sum_i \alpha_i y_i = 0$
  - C and margin (and also slack) are inversely related
    - $C \to 0 \Rightarrow \alpha_i \to 0 \Rightarrow \theta \to 0$ then you pay no attention to the constraints and $\theta \to 0$
  - If we add a constant to $K(x_i, x_j) \leftarrow K(x_i, x_j) + 1$, what happens to $\phi(x)$?
    - $\phi'(x) = \begin{bmatrix} \phi(x) \\ 1 \end{bmatrix}$
    - If we already have a θ_0 in the constraints, then that constant won't matter
      - $\frac{1}{2}\sum_i \sum_j \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j) + 1 = \frac{1}{2}\sum_i \sum_j \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j) + \frac{1}{2}(\sum_i \alpha_i y_i)^2 = \frac{1}{2}\sum_i \sum_j \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j)$

Linear regression

- Kernel formulations
- Linear regression as a statistical model $y = \theta\phi(x) + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$
  - We looked at MLE, MAP and…
  - … Bayesian estimation of θ: Given distribution on θ, and some training data, how do I adjust θ?
    - Translating Bayesian estimation into a Gaussian process, if we assume θ is distributed like Gaussian. The Guassian was like the dual of the Bayesian estimation
      - $C(x_i, x_j) = K(x_i, x_j) + \delta_{i,j}\sigma^2$
      - Expected to know how covariance function came about, and how it might behave if we change the noise:
        - A lot of noise means I will barely react to training data
  - Difference between MAP and Bayesian? MAP is just regularization. (WHAT?)
    - $\max \log P(y_i | x_i, \theta) + \log P(\theta)$
    - Bayesian estimation will return a full distribution over the parameters.
- Regularization, how does the solution change as you change the regularization?
- In Bayesian estimation:
  - $y = \int P(y \mid x, \theta) P(\theta \mid S_n) d\theta$

Gaussian process

- What they mean
- How to make prediction with a Gaussian process
- Fair question: Given Gaussian process, what is the predictor? What is its variance?

Alin Tomescu

6.867 Machine learning | Week 7, Thursday, October 17th, 2013| Lecture 12

Classification and regression trees

- Estimate them greedily by splitting the input space
- Bootstrap average with random forest
    - Expected to know what is random forest, what are the key parameters, how it might behave as a key of varying the parameters
- Boosting as a way of training ensembles
    - How to estimate $h_m(x) = \sum_i \alpha_i h(x; \theta_i)$
    - Loss functions
    - AdaBoost: you can solve α_m once you fixed the base learner.
    - $\min \sum_i Loss\big(y_i h_{m-1}(x) + \alpha_m y_i h(x_i, \theta_m)\big) \Rightarrow \frac{\partial}{\partial \alpha_m}\big(\sum_i Loss\big(y_i h_{m-1}(x) + \alpha_m y_i h(x_i, \theta_m)\big)\big) = 0$
    - We looked at boosting by looking at $\overrightarrow{h_{m-1}} = [\cdots h_{m-1}(x_i) \cdots]^T$
        - We search for a direction that minimizes the weighted error most
        - We decide how far to go along in that direction in order to minimize the loss
            - The weights will change to remove the advantage of that base learner, so that we don't re-pick it in a later step.

Generalization

- We would expect you to know that generalization depends on the size of the set of classifiers that's under consideration
- Regularization tries to squeeze the size of things that I have under consideration
    - That's why it helps to regularize, but it increases bias

Bias and variance, model selection

- Understand in general terms what bias and variance is, how are they impacted by choice of parameters.
- If I squeeze the set of classifiers, variance goes down, but the bias increases, because I have only a few choices.

$$E\left\{\left(\hat{y}(x) - y^*(x)\right)^2\right\} = E_x\left\{\left(E\{\hat{y}(x)\} - y^*(x)\right)^2\right\} + E\{Var(\hat{y}(x))\}$$