Notes: Midterm 2 Review

Note 1. Recall that your midterm will be worth 64 points. It will cover only the *statistical* concepts we have learned so far, not the *computational*. In particular, I will not ask you about:

1. PLA,

2. TEA,

3. loss functions (e.g. convex/surrogate/robust), or

4. optimization (e.g. gradient descent/stochastic gradient descent).

Your grading will be divided into 4 parts, each worth 16 points:

1. Response to initial question.

2. Knowledge of Chapter 1 (everything except PLA, including the between chapter 1-2 notes packet).

3. Knowledge of Chapter 2 (sections 2.1 and 2.2).

4. Knowledge of Chapter 3 (section 3.4 only).

Your initial question will have the following form. (Essentially every technical machine learning interview will have a question like this. Also, every real world machine learning project has to answer this type of question multiple times.)

Problem 1. At ACME corps, we have deployed a perceptron model with the polynomial feature map of degree 4. Our training dataset size is $N = 10^4$ and $d = 10^2$. We have measured our in sample error as 0.13 and our test set error is 0.42. What should we do to improve our performance?

Your answer should:

1. Be less than 5 minutes in total. (1-2 minutes is ideal.)

2. First state whether you are overfitting the data or underfitting the data.

3. Draw the model complexity curve.

4. Use the curve to explain how to correct the over/underfitting problem.

    (a) Suggest explanations about why the model is over/underfitting.
    (b) Suggest concrete changes to either the hypothesis class or feature maps.
    (c) Explicitly state how this affects the VC dimension.
    (d) Explicitly state how the VC dimension relates to the model complexity curve. (That is, why is the change from the previous part good?)

Note 2. Follow up questions will assess your knowledge on each of the chapters. I will (try to) not ask a question directly from the notes packet, and instead ask questions related to what you talk about in your initial answer.

Expect to:

1. Talk about each of the definitions from the definitions quizzes. You should be able to define all of these terms and talk about why they are important in context.

2. Walk through the steps for calculating the VC dimension of a hypothesis class.

Here are more example initial questions. Any possible combination of hypothesis classes, feature maps, dataset size, and in/out of sample errors is fair game.

Problem 2. At ACME corps, we have deployed a decision stump model with the random feature map of output degree 10. Our training dataset size is $N = 10^5$ and dimensions is $d = 10^2$. We have measured our in sample error as 0.14 and our test error is also 0.14. What should we do to improve our performance?

Problem 3. At ACME corps, we have deployed the concentric circles model. We first center out data points and then apply the polynomial feature map with degree 20. Our training dataset size is $N = 10^2$ and dimensions is $d = 10^2$. We have measured our in sample error as 0.30 and our test error is 0.31. What should we do to improve our performance?

Problem 4. At ACME corps, we have deployed the $\mathcal{H}_{\text{axis2}}$ model with the PCA feature map after centering and rescaling the data. Our training dataset size is $N = 10^2$ and dimensions is $d = 10^6$. We have measured our in sample error as 0.05 and our test error is 0.06. What should we do to improve our performance?

Problem 5. At ACME corps, we have deployed the perceptron model with two feature maps: First we apply the polynomial feature map of degree 7, then we apply PCA with output dimension 5. Our training dataset size is $N = 10^5$ and dimensions is $d = 10^2$. We have measured our in sample error as 0.25 and our test error is 0.24. What should we do to improve our performance?

Problem 6. At ACME corps, we have deployed the model of convex sets in two dimensions. We use PCA to reduce the dimensions down to 2 and then rescale and center our data. Our training dataset size is $N = 10^8$ and dimensions is $d = 10^1$. We have measured our in sample error as 0.15 and our test error is 0.42. What should we do to improve our performance?

Note 3. Practicing these problems with another student will count for the midterm1 extra credit.