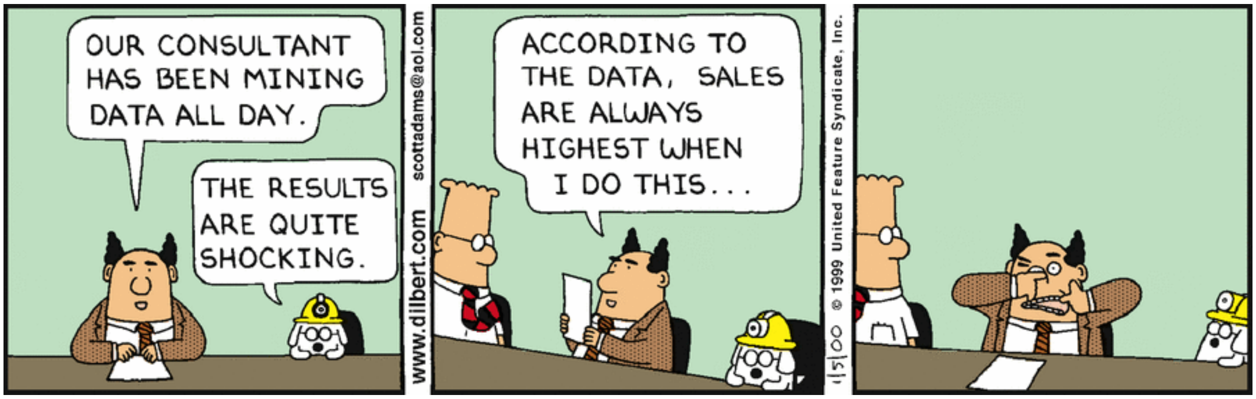
Notes: Statistical Learning Theory I


OUR CONSULTANT HAS BEEN MINING DATA ALL DAY.
THE RESULTS ARE QUITE SHOCKING.
scottadams@aol.com
www.dilbert.com
ACCORDING TO THE DATA, SALES ARE ALWAYS HIGHEST WHEN I DO THIS...
© 1999 United Feature Syndicate, Inc.
1/5/00

1 Pre-lecture Work

Data mining has a different philosophical approach to statistics than you were probably taught in your stats class. In particular, we do not try to falsify hypotheses; there are no p-values; and we do not try to find out how the world "really works." Instead, we're trying to find models that are "good enough" in the sense that they predict the future well. Data miners embrace the maxim that:

> *All models are wrong, but some models are useful.*[1]

In this section on Statistical Learning Theory, we are going to formalize what it means for a model to be "useful" and how you select the useful models from the non-useful ones. In particular, we are going to formalize the following claim:

> *Complex models require more data to train than simple models in order to predict the future well.*

There are many ways to formalize what it means for a model to be "more" and "less" complex, and in this class we will cover a simple complexity measure known as the *VC dimension*. In order to define the VC dimension, we must first introduce quite a bit of notation this week, and it's perfectly okay if this does not seem simple to you. Next week we will actually introduce the VC dimension and its applications.

Problem 1. (optional) The following two youtube videos review some basic statistics/machine learning terms.

`https://www.youtube.com/watch?v=Gv9_4yMHFhI`

`https://www.youtube.com/watch?v=EuBBz3bI-aA`

Define the following terms as presented in the videos.

1. training data

[1] `https://en.wikipedia.org/wiki/All_models_are_wrong`

2. testing data

3. bias-variance tradeoff

4. bias

5. variance

6. overfit

7. underfit

What is the purpose of the techniques *regularization*, *boosting*, and *bagging*?

Problem 2. (optional) The following youtube video reviews ordinary least squares (OLS) regression.

`https://www.youtube.com/watch?v=nk2CQITm_eo`

During lecture, we will be using OLS as an example, and so being familiar with the technique will help you understand lecture.

Problem 3. Read Chapter 2 of Shalev-Shwartz and Ben-David. Complete the following notes as you read.

1. Reproduce equations 2.1, 2.2 below.

   Note: Whenever I ask you to reproduce an equation/definition/theorem from the book, you should also provide definitions of all of the variables used in these equations.

2. Reproduce the definition of the realizability assumption (Definition 2.1).

3. Reproduce Corollary 2.3 below.

Problem 4. Read Chapter 3. Define the following terms:

1. PAC Learnability (Definition 3.1)

2. Corollary 3.2

   Notice that this is equivalent to Corollary 2.3, but we have "refactored" the corollary to rely on PAC learnability.

3. Equation 3.1
   Specifically highlight the difference between 3.1 and 2.1

4. Bayes optimal predictor $f_{\mathcal{D}}$

5. Agnostic PAC Learnability (Definition 3.3)
   Highlight the difference between the definition of agnostic PAC and plain PAC learnabillity.

6. Equation 3.3

7. Equation 3.4

8. 0-1 loss

9. square loss

10. Agnostic PAC Learnability for General Loss Functions (Definition 3.4)

    Again, highlight the difference between the definition of agnostic PAC learnability and agnostic PAC learnability for general loss functions.

Problem 5. Let $\mathcal{H}$ be a finite hypothesis class with $|\mathcal{H}| = 1000$. Use corollary 2.3 to calculate the number of data points needed to ensure that the empirical risk minimizer h_S has true risk less than 0.1 with 99% probability. That is,

$$L_{(\mathcal{D},f)}(h_S) \leq 0.1. \tag{1}$$

Problem 6. For each statement below, indicate whether the statement is true or false. If it is true, prove it; if false, explain why or provide a counter example.

1. It is possible for $L_S(h_S)$ to be less than $L_{\mathcal{D}}(h_S)$.

2. Let S' be a "test set" of m' data points sampled from $\mathcal{D}^{m'}$ and let h_S be a hypothesis trained on dataset S using any learning algorithm (e.g. possibly one other than the empirical risk minimizer). Then,

$$\mathbb{E}_{S'}\, L_{S'}(h_S) = L_{\mathcal{D}}(h_S). \tag{2}$$

The function $L_{S'}$ is defined analogously to the function L_S in Equation 2.2.

3. All finite hypothesis classes are PAC learnable.

4. There exists a hypothesis h that has lower error than the Bayes Optimal Predictor $f_{\mathcal{D}}$.

5. If the hypothesis class $\mathcal{H}$ satisfies the realizability assumption, then the Bayes Optimal Predictor $f_{\mathcal{D}}$ must be contained in $\mathcal{H}$.

6. If the Bayes Optimal Predictor $f_{\mathcal{D}}$ is contained in the hypothesis class $\mathcal{H}$, then $\mathcal{H}$ must satisfy the realizability assumption,

7. If the hypothesis class $\mathcal{H}$ satisfies the realizability assumption, then the Bayes Optimal Predictor $f_{\mathcal{D}}$ must have a true risk of zero. That is,

$$L_{\mathcal{D}}(f_{\mathcal{D}}) = 0. \tag{3}$$

8. If the Bayes Optimal Predictor $f_{\mathcal{D}}$ has a true risk of zero, then the hypothesis class $\mathcal{H}$ must satisfy the realizability assumption,

9. If there exists a hypothesis h in hypothesis class $\mathcal{H}$ such that

$$\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) = 0, \tag{4}$$

then $\mathcal{H}$ is PAC learnable.

10. If there exists a hypothesis h in hypothesis class $\mathcal{H}$ such that

$$\min_{h \in \mathcal{H}} L_{S}(h) = 0, \tag{5}$$

then $\mathcal{H}$ is PAC learnable.

11. Every hypothesis class that is agnostic PAC learnable is necessarily also PAC learnable.

12. Every hypothesis class that is PAC learnable is necessarily also agnostic PAC learnable.

13. Let S be a dataset sampled from $\mathcal{D}^m$. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two hypothesis classes, with $\mathcal{H}_1 \supset \mathcal{H}_2$. Let $\hat{h}_1$ and $\hat{h}_2$ be corresponding empirical risk minimizers trained on S. That is,

$$\hat{h}_1 = \underset{h \in \mathcal{H}_1}{\arg\min}\, L_S(h) \qquad \text{and} \qquad \hat{h}_2 = \underset{h \in \mathcal{H}_2}{\arg\min}\, L_S(h). \tag{6}$$

(a) The emperical risk of $\hat{h}_1$ is guaranteed to be less than the emperical risk of $\hat{h}_2$. That is,

$$L_S(\hat{h}_1) \leq L_S(\hat{h}_2). \tag{7}$$

(b) The true risk of $\hat{h}_1$ is guaranteed to be less than the true risk of $\hat{h}_2$.

$$L_{\mathcal{D}}(\hat{h}_1) \leq L_{\mathcal{D}}(\hat{h}_2). \tag{8}$$

14. Let hypothesis class $\mathcal{H}$ be PAC learnable with sample complexity

$$m_{\mathcal{H}} = \frac{10^{1000} + \log(1/\delta)}{\epsilon}. \tag{9}$$

Then there does not exist a distribution $\mathcal{D}$ such that if $S \sim \mathcal{D}^m$ and h_S is the empirical risk minimizer trained on S, then the generalization error is

$$L_{\mathcal{D}}(h_S) \leq \frac{\log(1/\delta)}{m_{\mathcal{H}}}. \tag{10}$$

2 Lecture

Problem 7. Prove or disprove the following claims. Let $\mathcal{H}$ be a finite hypothesis class. Then the sample complexity of $\mathcal{H}$ is bounded by

$$m_{\mathcal{H}} = \Omega(1) \qquad \text{and} \qquad m_{\mathcal{H}} = O\Bigg(\log(|\mathcal{H}|) \Bigg). \tag{11}$$

Problem 8. Let $\mathcal{H}$ be a hypothesis class for a binary classification task. Suppose that $\mathcal{H}$ is PAC learnable and its sample complexity is given by $m_{\mathcal{H}}(\cdot, \cdot)$.

1. Show that $m_{\mathcal{H}}$ is monotonically nonincreasing in ϵ. That is, show that given $\delta \in (0, 1)$, and given $0 < \epsilon_1 \leq \epsilon_2 < 1$, we have that $m_{\mathcal{H}}(\epsilon_1, \delta) \geq m_{\mathcal{H}}(\epsilon_2, \delta)$.

2. Show that $m_{\mathcal{H}}$ is monotonically nonincreasing in δ. That is, show that given $\epsilon \in (0, 1)$, and given $0 < \delta_1 \leq \delta_2 < 1$, we have that $m_{\mathcal{H}}(\delta_1, \epsilon) \geq m_{\mathcal{H}}(\delta_2, \epsilon)$.

Problem 9. This problem concerns ordinary least squares (OLS) regression.

1. Define the OLS model.

2. Define the hypothesis class.

3. Which loss function does OLS regression use?

4. Which PAC Learning definition(s) make sense for the OLS problem.

5. Provide a closed form solution for the ERM.

   Hint: for calculating matrix derivatives, you can use `http://www.matrixcalculus.org/`

6. What is the runtime of computing the ERM for OLS?

7. Assume that each parameter is implemented using a 64-bit floating point number. Bound the sample complexity of OLS.

8. Assume that each parameter is implemented using an b-bit floating point number. Bound the sample complexity of OLS.

9. A d dimensional OLS model is trained on a dataset S. The training error is 3.5 and the testing error is 12.4. Is the model over-fitting or under-fitting?

   How should we adjust the dimensionality d of the model in order to improve performance on the test set?

10. A d dimensional OLS model is trained on a dataset S. The training error is 12.4 and the testing error is 3.5. Is the model over-fitting or under-fitting?

    How should we adjust the dimensionality d of the model in order to improve performance on the test set?