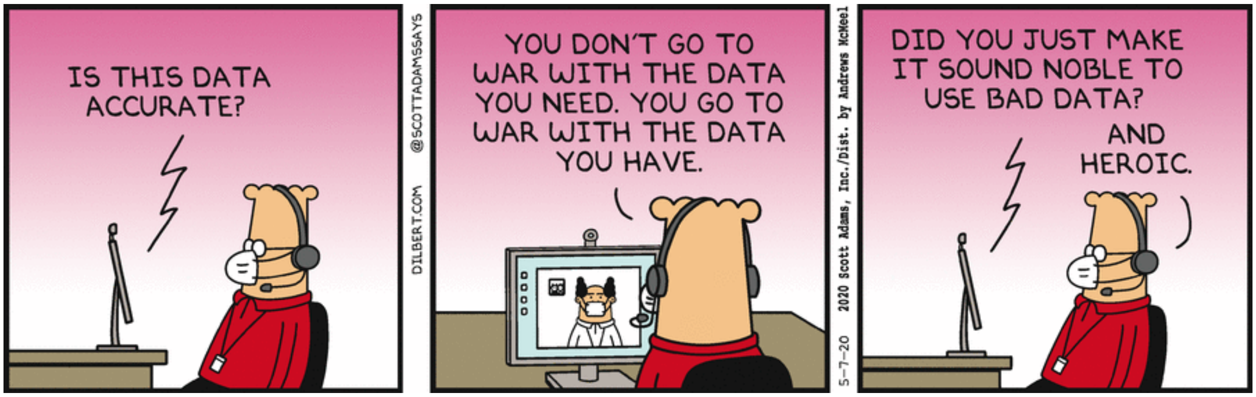
Notes: Statistical Learning Theory II


IS THIS DATA ACCURATE?
DILBERT.COM @SCOTTADAMSSAYS
YOU DON'T GO TO WAR WITH THE DATA YOU NEED. YOU GO TO WAR WITH THE DATA YOU HAVE.
5-7-20 2020 Scott Adams, Inc./Dist. by Andrews McMeel
DID YOU JUST MAKE IT SOUND NOBLE TO USE BAD DATA?
AND HEROIC.

1 Pre-lecture Work

Problem 1. (optional) Chapter 4 of Shalev-Shwartz and Ben-David covers a concept called *uniform convergence*. This is a mathematical tool that was historically used before the discovery of the VC-dimension, and is currently used in situations where the VC-dimension is not applicable. We will not cover the concept in this class, but if you are particularly interested in machine learning theory, then I recommend reading section 4.1 from this chapter. (It's only 2 pages.) Then in Section 4.2, the authors generalize Corollary 2.3 (finite hypothesis classes are PAC learnable) to the agnostic setting.

Problem 2. Read Chapter 5 of Shalev-Shwartz and Ben-David. (You may skip section 5.1, which formally defines the *No Free Lunch Theorem*.) Complete the following notes as you read.

1. Equation (5.7)

Problem 3. Read Chapter 6 of Shalev-Shwartz and Ben-David. (You may skip Section 6.5, which is only concerned with proofs.) Complete the following notes as you read.

1. Restriction of $\mathcal{H}$ to C (Definition 6.2)

2. Shattering (Definition 6.3)

3. VC-Dimension (Definition 6.5)

4. Theorem 6.6

5. The Fundamental Theorem of Statistical Learning (Theorem 6.7). You may ignore result 1 about uniform convergence (we're not covering uniform convergence in this class). You only need to know results 2-6.

6. The Fundamental Theorem of Statistical Learning - Quantitative Version (Theorem 6.8). You may ignore result 1 about uniform convergence (we're not covering uniform convergence in this class). You only need to know results 2 and 3 about PAC and agnostic PAC learnability.

Problem 4. For each hypothesis class below, formally define the hypothesis class and state its VC-dimension. You can find all of these answers in Section 6.3. You do not need to provide the proof of the VC-dimension below.

1. Threshold functions

2. Intervals

3. Axis Aligned Rectangles

4. Finite Classes

Problem 5. Prove or disprove the following statements. Note that all of the proofs/disproofs follow immediately from the definitions above, and that is why they are included in this section. You do not have to complete all of these problems before the start of lecture. We will discuss some of these problems during lecture, but I recommend you solve as many as you can on your own before lecture.

1. The following equation always holds:

$$L_{\mathcal{D}}(h_S) - \epsilon_{\text{est}} = \epsilon_{\text{app}} \tag{1}$$

2. The following equation always holds:

$$L_{\mathcal{D}}(h_S) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) = \epsilon_{\text{est}} \tag{2}$$

3. The following equation always holds:

$$\epsilon_{\text{app}} > 0 \tag{3}$$

4. The following equation always holds:

$$\epsilon_{\text{est}} > 0 \tag{4}$$

5. The following equation always holds:

$$\epsilon_{\text{est}} < \epsilon_{\text{app}} \tag{5}$$

6. As the number of data points in the training set increases, the approximation error decreases.

7. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be hypothesis classes where $\mathcal{H}_1 \subset \mathcal{H}_2$. Then, the approximation error of $\mathcal{H}_1$ is greater than or equal to the approximation error of $\mathcal{H}_2$.

8. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be hypothesis classes where $\mathcal{H}_1 \subset \mathcal{H}_2$. Then, the estimation error of $\mathcal{H}_1$ is less than or equal to the estimation error of $\mathcal{H}_2$.

9. A model with a high approximation error and a low estimation error is underfitting.

10. A model with a low approximation error and a high estimation error is overfitting.

11. If $\mathrm{VCdim}(\mathcal{H})$ is finite, then $\epsilon_{\mathrm{app}} = 0$.

12. If $\mathrm{VCdim}(\mathcal{H})$ is infinite, then $\epsilon_{\mathrm{app}} = 0$.

13. If $\mathrm{VCdim}(\mathcal{H})$ is finite, then the Bayes optimal predictor $f_{\mathcal{D}} \in \mathcal{H}$.

14. If $\mathrm{VCdim}(\mathcal{H})$ is infinite, then the Bayes optimal predictor $f_{\mathcal{D}} \notin \mathcal{H}$.

15. If $\mathrm{VCdim}(\mathcal{H})$ is infinite, then $L_{\mathcal{D}}(h) > 0$ for all distributions $\mathcal{D}$, datasets $S \sim \mathcal{D}^m$, and $h \in \mathcal{H}$.

16. If VCdim($\mathcal{H}$) is infinite, then $L_S(h_S) = 0$ for all distributions $\mathcal{D}$, datasets $S \sim \mathcal{D}^m$, and any ERM $h_S \in \mathcal{H}$.

17. If VCdim($\mathcal{H}$) is finite, then $\mathcal{H}$ is PAC learnable.

18. If $\mathcal{H}$ is agnostic PAC learnable, then it is also PAC learnable.

19. For every two hypotheses class $\mathcal{H}_1$ and $\mathcal{H}_2$, if $\mathcal{H}_1 \subset \mathcal{H}_2$, than $\text{VCdim}(\mathcal{H}_1) \leq \text{VCdim}(\mathcal{H}_2)$.

20. For every two hypotheses class $\mathcal{H}_1$ and $\mathcal{H}_2$, if $\text{VCdim}(\mathcal{H}_1) = \text{VCdim}(\mathcal{H}_2)$, then $\mathcal{H}_1 = \mathcal{H}_2$.

21. The ordinary least squares (OLS) hypothesis class discussed in the previous lecture notes has a finite VC dimension.

2 Lecture

Problem 6. Combine Theorem 6.8 and the definition of (agnostic) PAC learnability to bound the generalization error of a hypothesis class based on its VC-dimension.

Problem 7. Linear models are one of the main tools in data mining. Chapter 9 in Shalev-Swartz and Ben-David discuss linear predictors in significantly more detail than we need for this class. In this problem, we will review all the relevant concepts.

1. Define the hypothesis class of halfspaces.

2. What is separability?

3. What is the VC-Dimension?

4. What is the computational complexity of computing the ERM in the separable and agnostic cases?

Problem 8. Kernel functions are tools that let us manipulate the VC-dimension of hypothesis classes. They also have nice computational properties, but in this problem we are only concerned with their statistical properties.

1. Define the polynomial kernel.

2. What is the VC-dimension of halspaces with the polynomial kernel?

3. When would we use the polynomial kernel?

4. Define the random projection kernel.

5. What is the VC-dimension of halspaces with the random projection kernel?

6. When would we use the random projection kernel?