Generalization Bounds

CMPUT 296: Basics of Machine Learning

Textbook Ch.12

Logistics

- Midterm **spot checks** have been scheduled
- **Thought questions #4** due **Thursday (Nov 26)**
  - Anything between logistic regression and generalization bounds is fair game
- **Last class** will be on **Tuesday (Dec 1)**

Outline

1. Recap & Logistics
2. Empirical Error
3. Measuring Hypothesis Class Size
4. Generalization Bounds

Recap: Optimal Prediction

Suppose we know the true joint distribution $p(\mathbf{x}, y)$, and we want to use it to make predictions in a classification problem.

The **optimal classification predictor** makes the **best** use of this function.

As with the optimal estimator, we measure the quality of a predictor $f(\mathbf{x})$ by its **expected cost** $\mathbb{E}[C(f)]$. The optimal predictor **minimizes** $\mathbb{E}[C(f)]$.

$$\mathbb{E}[C(f)] = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} \text{cost}\left(f(\mathbf{x}), y\right) p(\mathbf{x}, y)\, d\mathbf{x},$$

where $\text{cost}(\hat{y}, y)$ is the cost for predicting $\hat{y}$ when the true value is y, and $C(f) = \text{cost}\left(f(X), Y\right)$ is a random variable.

Empirical Error Estimation

- We can't actually measure **generalization error**

$$C(f) = \mathbb{E}\left[\text{cost}(f(\mathbf{X}), Y)\right]$$

- But we can **estimate** it with the **empirical error** on a dataset $\mathscr{D} = \{(\mathbf{x}_i, y_i) \mid 1 \leq i \leq n\}$:

$$\hat{C}(f) = \frac{1}{n}\sum_{i=1}^{n} \text{cost}\left(f(\mathbf{x}_i), y_i\right)$$

- The empirical error is an **unbiased estimator** for the generalization error:

$$\mathbb{E}\left[\hat{C}(f)\right] = C(f)$$

Question: Doesn't this conflict with "don't use the training set to estimate generalization error?"

Empirical Risk Minimization

Since $\hat{C}(f)$ is a consistent estimator for our target $C(f)$, one strategy is to minimize this estimator directly. This is called **empirical risk minimization**:

$$\hat{f}_{\text{ERM}} = \arg\min_{f \in \mathcal{F}} \hat{C}(f)$$

E.g., ordinary least squares minimizes the empirical squared cost over the linear hypothesis class $\mathcal{F} = \left\{ f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} \mid \mathbf{w} \in \mathbb{R}^d \right\}$:

$$\mathbf{w}_{\text{OLS}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \mathbf{w}^\top \mathbf{x} \right)^2$$

Recap: Bias vs. Variance

$$\mathbb{E}\left[\left(f_{\mathscr{D}}(X) - f^*(X)\right)^2\right] = \left(\mathbb{E}\left[f_{\mathscr{D}}(X)\right] - f^*(X)\right)^2 + \text{Var}\left[f_{\mathscr{D}}(X)\right]$$

- We can decompose the **reducible error** into bias and variance of the **predictions**
- Note that $f^*(X)$ is the **optimal predictor**; it **need not** be part of our **hypothesis class**
- $f_{\mathscr{D}}(X)$ is the **predictor** that will be chosen from our **hypothesis class** based on the dataset $\mathscr{D}$ (so when we treat $\mathscr{D}$ as a random variable, $f_{\mathscr{D}}$ is also random)
- Choosing a different hypothesis class **can change** both the bias and variance of $f_{\mathscr{D}}$
  - "Bigger" hypothesis class: More variance, because it can fit a dataset in more ways
  - "Smaller" hypothesis class: More bias, because it can only fit some functions

Bounding vs. Cross-Validation

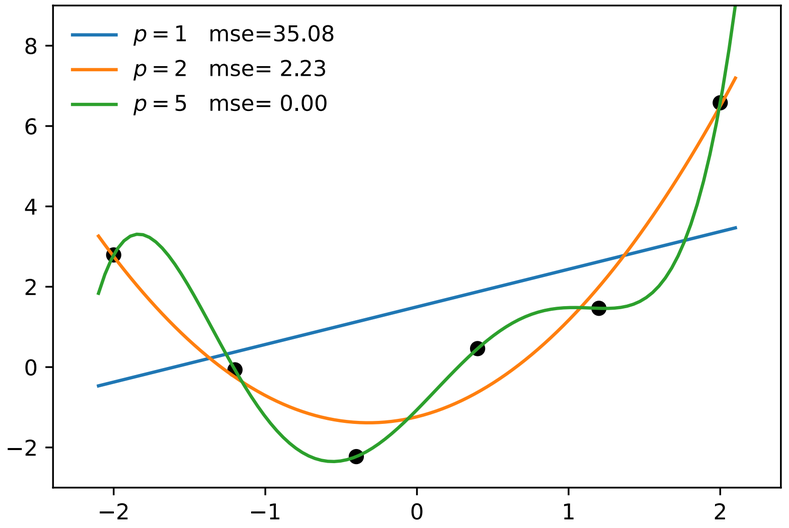
- How can you choose a hypothesis class?
- **Lecture 15 approach:** Cross-validation to choose a hypothesis class:

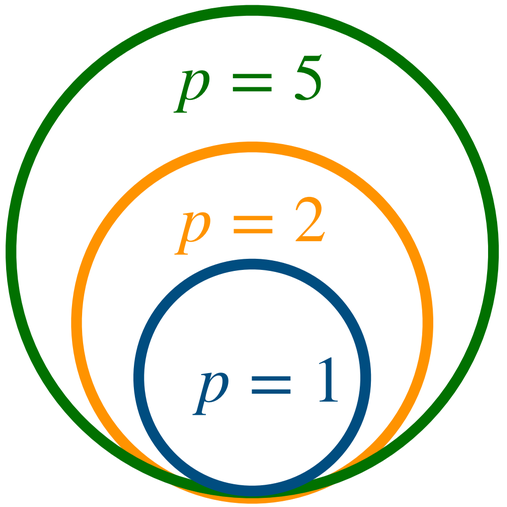
1. Perform k-fold cross-validation for p-degree polynomial regression for all $1 \leq p \leq P$
2. Let p^* be the p that minimizes the estimated generalization error
3. Fit a p^*-degree polynomial on the full training dataset

- **This lecture:** compute an **upper bound** on the error of a predictor
  - Then choose hypothesis class to minimize that upper bound

Quantifying the "Size" of a Hypothesis Class

- We know that the class of quadratic functions is "bigger" than the class of linear functions, because it is a **superset** of the linear functions
- But can we put a number on this difference?
- This is *not* about counting the **number of hypotheses** contained in the class (**why?**)
  - (They are both infinite!)
- How much more expressive is the class of quadratic hypotheses than the class of linear hypotheses?





Empirical Rademacher Complexity

The **empirical Rademacher complexity** of $\mathscr{F}$ with respect to $\mathscr{D}$ is

$$\hat{R}_{\mathscr{D}}(\mathscr{F}) = \mathbb{E}\left[\max_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i)\right]$$

Question: What is the expectation taken over?

where

- $\mathscr{D} = \{(\mathbf{x}_i, y_i) \mid 1 \leq i \leq n\}$ is a dataset
- $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ is a vector of n random variables, with $\sigma_i \overset{i.i.d}{\sim} \text{Uniform}\{-1, +1\}$ **("Rademacher variables")**
- $\mathscr{F}$ is a hypothesis class

- Intuitively, the Rademacher complexity measures how well we can correlate with **random noise** by choosing a hypothesis from $\mathscr{F}$.
- More complex hypothesis classes can better correlate with random noise, because they have more functions to choose from.

Rademacher Complexity

The **Rademacher complexity** of a hypothesis class $\mathscr{F}$ is the expected **empirical Rademacher complexity** over all datasets $\mathscr{D}$ (of size n):

$$R_n(\mathscr{F}) = \mathbb{E}\left[\hat{R}_{\mathscr{D}}(\mathscr{F})\right]$$

Question: What is the expectation taken over?

- The empirical Rademacher complexity is with respect to a single, fixed dataset
- The Rademacher complexity is with respect to the **distribution** of datasets
  - Note that the Rademacher complexity of a hypothesis class can differ depending on the data distribution!

Example: $(n-1)$-Degree Polynomial

$$\hat{R}_{\mathcal{D}}(\mathcal{F}) = \mathbb{E}\left[\max_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i)\right]$$

- Let $\mathcal{F} = \{f_{\mathbf{w}}(x) = w_0 + w_1 x \mid w_0, w_1 \in [-1,1]\}$
- Let $\mathcal{D} = \{(1, 4), (2, -5)\}$
- What is $R_{\mathcal{D}}(\mathcal{F})$?

Caveat: We will ignore y_i
This is usually handled in a different way

σ_1	σ_1	w0	w1	∑f(x) / 2
-1	-1	1	-1	2.5
-1	+1	1	1	1.5
+1	-1	1	-1	2.5
+1	+1	1	1	1.5

  - We need to be able to make $f(x_i)$ very positive or very negative, depending on σ
  - $R_{\mathcal{D}}(\mathcal{F}) = 2$ for the given dataset and hypothesis class
- Notice that we are now assuming that the weights are **bounded**
  - **Question:** What would the Rademacher complexity be for $\mathcal{F} = \{f_{\mathbf{w}}(x) = w_0 + w_1 x \mid \mathbf{w} \in \mathbb{R}^2\}$?

Generalization Bound for a Hypothesis Class

Theorem:
Let $\mathcal{F}$ be a family of binary classification functions taking values in $\{-1, +1\}$. Then for every $f \in \mathcal{F}$, and every $\delta > 0$,

$$C(f) \leq \hat{C}(f) + R_n(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2n}}$$

with probability $1 - \delta$.

Questions: As n grows,

1. What happens to $R_n(\mathcal{F})$?
2. What happens to $\sqrt{\frac{\log 1/\delta}{2n}}$?
3. What happens to $|C(f) - \hat{C}(f)|$?

- *Idea:* Rather than optimizing $\hat{C}(f)$, optimize the whole RHS
- **Question:** What good would that do?

Summary

- "Larger" hypothesis classes have smaller bias, but larger variance
- Generalization error decomposes into bias and variance terms
    - We might prefer a "smaller" hypothesis class if we could reduce variance enough to make up for the increased bias
- **Empirical Risk Minimization:** directly optimize the loss on the training set
- **Rademacher complexity:** Measures the "size" of a hypothesis class by its ability to fit random noise
- We can upper bound generalization error by the sum of empirical cost, Rademacher complexity of the hypothesis class, and another term
    - This can help guide us in our decisions about which hypothesis class to use