Evaluation of Models & Hypothesis Testing

CMPUT 296: Basics of Machine Learning

Textbook §8.3

Logistics

- Quiz and Thought Questions #2 have been marked
  - See eclass for marks and comments
  - Question 6 (derive optimal predictor for a given cost function) seemed to give people particular trouble
- **Assignment #2** is due on **Thursday (Oct 22)**
- **Midterm exam** is **next Thursday (Oct 29)**

Recap: Generalization & Overfitting

- Our goal is to minimize **generalization error**: expected cost with respect to the **underlying distribution**
- But we only have access to **empirical error**: average cost on a dataset
- The empirical error of a model on its **training data** is a **biased, over-optimistic estimate** of generalization error
- Using an **overly complex** model leads to **overfitting**:
  High training performance at the expense of generalization performance
  - **Underfitting** comes from using an **overly simple** model
- A **held-out test set** gives an unbiased estimate of generalization error
  - But you can only use it **once**!
  - Alternatives: k-fold cross-validation; bootstrap resampling

Outline

1. Recap & Logistics
2. Confidence Intervals
3. Hypothesis Tests

Probabilistic Comparison

- We can use a **test set** to obtain m **samples** of generalization error
  - (or k-fold cross-validation, or bootstrap resampling, or...)
- We can estimate the **generalization error** of models f_1 and f_2 by the **empirical costs**

$$\hat{C}_1 = \frac{1}{m}\sum_{i=1}^{m} c_i(f_1) \text{ and } \hat{C}_2 = \frac{1}{m}\sum_{i=1}^{m} c_i(f_2), \qquad \text{where } c_i(f) = \mathrm{cost}(f(\mathbf{x}_i), y_i)$$

Questions

1. Suppose that $\hat{C}_1 < \hat{C}_2$. Is f_1 a **better** model than f_2?
2. If $\hat{C}_1 < \hat{C}_2$, with what **probability** is f_1 a better model than f_2?

Confidence Intervals

- One approach is to make claims of the form

$$\Pr\left[\left|\hat{C} - \mathbb{E}[C]\right| \leq \epsilon\right] \geq 1 - \delta$$

- i.e., compute **a $(1-\delta)$ confidence interval** $[\hat{C} - \epsilon, \hat{C} + \epsilon]$
- Suppose that we assume that our error is **bounded** $a \leq c_i(f) \leq b \quad \forall f, i$
  - **Question:** Is that a plausible assumption?
- **Question:** How could we use that assumption to find a confidence interval?
- We can compute confidence intervals using concentration inequalities such as Hoeffdings's Inequality or Chebyshev's inequality
  - However, we typically make a distributional assumption instead (**why?**)

Gaussian Confidence Interval

- Suppose that we know that we assume that our errors $c_i(f)$ have a **Gaussian distribution**
    - **Question:** Is that a plausible assumption?
- If the errors have a Gaussian distribution, then we can find a $95\ \%$ confidence interval as simply $[\hat{C} - 1.96\sigma/\sqrt{m}, \hat{C} + 1.96\sigma/\sqrt{m}]$
    - More generally: $[\hat{C} - z_{\delta/2}\sigma/\sqrt{m}, \hat{C} + z_{\delta}\sigma/\sqrt{m}]$ for $z_{\delta/2} = \Phi^{-1}(\delta/2)$
- This will tend to give much tighter bounds than concentration inequalities
- **Question:** What is the problem with this approach?
- **Question:** Is it plausible to assume that we know σ?

Student's t-Distribution

- As an alternative, we can assume that the errors have a Student's t-distribution with $m - 1$ **degrees of freedom**
- A $1 - \delta$ confidence interval for a sample of m costs, assuming that each cost is normally distributed, is given by $[\hat{C} - \epsilon, \hat{C} + \epsilon]$, where

$$\epsilon = t_{\delta/2,m-1}\frac{S_m}{\sqrt{m}} \text{ and } S_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}(c_i(f) - \hat{C})^2$$

- $t_{\delta,m-1}$ depends on δ (as with Gaussian CI); also now depends on m
  - as $m \to \infty$, $t_{\delta/2,m-1} \to z_{\delta/2}$ (i.e., $t_{\delta/2,m-1} \to \Phi^{-1}(\delta/2)$)
- However, this expression does not depend on the unknown true variance σ
  - S_m^2 is the "Bessel corrected" variance estimator (often called the **sample variance**)

Comparing Two Models

- Suppose that we have $(1-\delta)$ confidence intervals for the generalization error of models f_1 and f_2: $[\hat{C}_1 - \epsilon_1, \hat{C}_1 + \epsilon_1]$ and $[\hat{C}_2 - \epsilon_2, \hat{C}_2 + \epsilon_2]$
- If $\hat{C}_1 + \epsilon_1 < \hat{C}_2 - \epsilon_2$, then we can say that f_1 is **statistically significantly** better than f_2 with confidence level δ:
- If $C_1 > C_2$, then at least one of the following must be true:

either $C_1 > \hat{C}_1 + \epsilon_1$ or $C_2 < \hat{C}_2 - \epsilon_2$

Union bound:

$\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$

- By the **union bound**:

$$\Pr\left[(C_1 > \hat{C}_1 + \epsilon_1) \vee (C_2 < \hat{C}_2 - \epsilon_2)\right] \leq \Pr\left[(C_1 > \hat{C}_1 + \epsilon_1)\right] + \Pr\left[(C_2 < \hat{C}_2 - \epsilon_2)\right] = \frac{\delta}{2} + \frac{\delta}{2} = \delta$$

Ranking Models

- Suppose we just want to rank two models, rather than quantifying their exact generalization error
- For a randomly-selected datapoint $(\mathbf{X}, Y)$, let

$$W = \begin{cases} 1 & \text{if } \operatorname{cost}\left(f_1(\mathbf{X}), Y\right) < \operatorname{cost}\left(f_2(\mathbf{X}), Y\right) \\ 0 & \text{otherwise.} \end{cases}$$

- The test set consists of m observations $W_1, \ldots, W_m \overset{i.i.d}{\sim} W$
- Let k be the number of "wins" (i.e., $w_i = 1$)
- Let $\beta = \Pr(W_i = 1)$

Question:

If f_1 is better than f_2, then what is β?

Hypothesis Test:
Binomial Counting Test

We want to do a **hypothesis test:**

$$H_0 : \beta = \frac{1}{2} \quad \text{vs} \quad H_1 : \beta > \frac{1}{2}$$

1. We compute the probability $p = \Pr\left[\sum_{i=1}^{m} W_i \geq k\right]$ of seeing at least k "wins", under the assumption that H_0 (the **null hypothesis**) is **true**
2. If $p < \alpha$, then we **reject** the null hypothesis with significance level of α
   - α is pretty arbitrary, but typically $\alpha \in \{0.01, 0.05, 0.10\}$

Hypothesis Test: Binomial Counting Test

$$\Pr(W_1 = w_1, \ldots, W_m = w_m) = \prod_{i=1}^{m} \left(w_i \beta + (1 - w_i)(1 - \beta) \right) = \beta^k (1 - \beta)^{m-k}$$

$$\Pr \left[\sum_{i=1}^{m} W_i = k \right] = \binom{m}{k} \beta^k (1 - \beta)^{m-k}$$

$$p = \Pr \left[\sum_{i=1}^{m} W_i \geq k \right] = \sum_{j=k}^{m} \Pr \left[\sum_{i=1}^{m} W_i = j \right] = \sum_{j=k}^{m} \binom{m}{j} \beta^j (1 - \beta)^{m-j}$$

So when $\sum_{j=k}^{m} \binom{m}{j} \left(\frac{1}{2} \right)^m < 0.05$, we can conclude that f_1 is significantly better than f_2, with $\alpha = 0.05$.

Hypothesis Tests: Paired t-Test

- Consider the dataset to be m observations of differences in cost: $c_i(f_1) - c_i(f_2)$
- If errors are distributed normally, then so are the differences
  - We don't know the variance, so use a t-distribution instead of Gaussian
- If the models are equally good, then expected value for each difference is 0
- Null hypothesis: expected value of difference is 0
- p-value: the probability that empirical average difference will be at least as large as $\bar{d} = \frac{1}{m}\sum_{i=1}^{m} d_i = \frac{1}{m}\sum_{i=1}^{m} c_i(f_1) - c_i(f_2)$

Which Test to Use?

- Each of these two tests makes **parametric assumptions**
- **Paired t-test:** Paired errors are i.i.d. normally distributed
  - **Question:** When might this assumption fail to hold?
- **Binomial counting test:** Compared values are in $\{0,1\}$
  - **Question:** When might this assumption fail to hold?
- Factors to consider:
  1. Applicability of the **assumptions**
  2. **Power** of the test: Probability of rejecting null when null is false
     - Confidence intervals are a low-power test

Summary

- We will often want to **compare** the generalization errors of two models
  - But we can't actually observe the generalization errors directly
- If the $(1 - \delta)$ **confidence intervals** for the two models do not overlap, then we say that one model has **statistically significantly** better generalization error than the other, with **confidence level** δ
- More powerful: paired **hypothesis test**, e.g.:
  - Binomial counting test
  - Paired t-test
- p**-value:** Probability of seeing our dataset given that null hypothesis is true
  - **Null hypothesis:** Both models have **equal errors**