Evaluating Performance

Which do you prefer and why?
Evaluating Performance

Which do you prefer and why?

Performance of a Fixed Hypothesis

(HTF 7.1–7.4, JWHT 2.2, 5)
• Define the loss (error) of the hypothesis on an example \((x, y)\) as

\[L(h(x), y)\]

• Suppose \((X, Y)\) is a vector-valued random variable. Then what is

\[L(h(X), Y)\]

**Performance of a Fixed Hypothesis**

• Given a model \(h\), (which could have come from anywhere), its *generalization error* is:

\[E[L(h(X), Y)]\]

• Given a set of data points \((x_i, y_i)\) that are realizations of \((X, Y)\), we can compute the *empirical error*

\[\bar{\ell}_{h,n} = \frac{1}{n} \sum_{i=1}^{n} L(h(x_i), y_i)\]

• What is \(\bar{\ell}_{h,n}\)?

**Generalization error of hypotheses from last day**

![Graph showing the relationship between the degree of polynomial and the log of the generalization error.](image)
Sample Mean

- Given a dataset (collection of realizations) $x_1, x_2, ..., x_n$ of $X$, the sample mean is:

$$\bar{x}_n = \frac{1}{n} \sum_i x_i$$

Given a dataset, $\bar{x}_n$ is a fixed number. We use $\bar{X}_n$ to denote the random variable corresponding to the sample mean computed from a randomly drawn dataset of size $n$.

Datasets and sample means

Datasets of size $n = 15$, sample means plotted in red.
Statistics, Parameters, and Estimation

- A **statistic** is any summary of a dataset. (E.g. $\bar{x}_n$, sample median.) A statistic is the result of a **function** applied to a dataset.

- A **parameter** is any summary of the distribution of a random variable. (E.g. $\mu_X$, median.) A parameter is the result of a **function** applied to a distribution.

- **Estimation** uses a **statistic** (e.g. $\bar{x}_n$) to estimate a **parameter** (e.g. $\mu_X$) of the **distribution** of a random variable.
  - **Estimate**: value obtained from a specific dataset
  - **Estimator**: function (e.g. sum, divide by n) used to compute the estimate
  - **Estimand**: parameter of interest

Sampling Distributions

(AoS, p.61, q.19)

Given an estimate, how good is it?
The distribution of an estimator is called its \textit{sampling distribution}.

\textbf{Bias}

\textit{(AoS, p.90)}

- The \textbf{expected difference} between estimator and parameter. For example,
  \[ E[\bar{X}_n - \mu_X] \]

  - If 0, estimator is \textbf{unbiased}.
  - Sometimes, $\bar{x}_n > \mu_X$, sometimes $\bar{x}_n < \mu_X$, but the long run average of these differences will be zero.

\textbf{Variance}

- The \textbf{expected squared difference} between estimator and its mean
  \[ E[(\bar{X}_n - E[\bar{X}_n])^2] \]

  - Positive for all interesting estimators.
  - For an unbiased estimator
    \[ E[(\bar{X}_n - \mu_X)^2] \]

  - Sometimes, $\bar{x}_n > \mu_X$, sometimes $\bar{x}_n < \mu_X$, but the \textit{squared differences} are all positive and do not cancel out.
Normal (Gaussian) Distribution

\[ f_X(x) = \frac{1}{\sigma_X \sqrt{2\pi}} e^{\frac{- (x-\mu_X)^2}{2\sigma_X^2}} \]

Normal distribution is defined by two parameters: \( \mu_X, \sigma_X^2 \).

The normal distribution is special (among other reasons) because *many estimators have approximately normal sampling distributions* or have sampling distributions that are closely related to the normal.

For an estimator like \( \tilde{X}_n \), if we know \( \mu_{\tilde{X}_n} \) and \( \sigma_{\tilde{X}_n}^2 \), then we can say a lot about how good it is.

Central Limit Theorem

\( \text{(AoS, p.77)} \)

- Informally: The sampling distribution of \( \tilde{X}_n \) is approximately normal if \( n \) is big enough.
- More formally, for \( X \) with finite variance:

\[ F_{\tilde{X}_n}(\tilde{x}) \approx \int_{-\infty}^{\tilde{x}} \frac{1}{\sigma_n \sqrt{2\pi}} e^{\frac{- (x-\mu\tilde{X}_n)^2}{2\sigma_n^2}} \]

where

\[ \sigma_n^2 = \frac{\sigma^2}{\sqrt{n}} \]

is called the *standard error* and \( \sigma^2 \) is the variance of \( X \).

Who cares?

- Eruptions dataset has \( n = 272 \) observations.
- Our estimate of the mean of eruption times is \( \tilde{x}_{272} = 3.4877831 \).
- What is the probability of observing an \( \tilde{x}_{272} \) that is within 10 seconds of the true mean?

Who cares?

By the C.L.T.,

\[ \Pr(-0.17 \leq \tilde{X}_{272} - \mu_X \leq 0.17) = \int_{x=-0.17}^{0.17} \frac{1}{\sqrt{2\pi}\sigma_n} e^{\frac{- (x-\mu\tilde{X}_n)^2}{2\sigma_n^2}} \]

\[ = 0.986 \]

Note! I estimated \( \sigma_X \) here. (Look up “t-test” for details.)
\[
\int_{x=-0.17}^{0.17} \frac{1}{\sqrt{2\pi\sigma_n}} e^{-\frac{(x-\mu_X)^2}{2\sigma_n^2}} = 0.986
\]

**Confidence Intervals**

(AoS, p.92)

- Typically, we specify *confidence* given by \(1 - \alpha\).
- Use the sampling distribution to get an *interval that traps the parameter (estimand) with probability* \(1 - \alpha\).
- 95% C.I. for eruption mean is \((3.35, 3.62)\)
95% Confidence Region
<table>
<thead>
<tr>
<th>-2</th>
<th>0</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>-2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>-2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>-2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>-2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>-2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>-2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>-2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>-2</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

What a Confidence Interval Means
Effect of $n$ on width

Performance Evaluation - Test Sets

Training error underestimates generalization error. It is a biased estimator.

- If you really want a good estimate of generalization error, you need to hold out a separate test set of data not used for training.

- Possibly of size $n = (1.96)^2 \frac{\sigma_L^2}{d^2}$ where $\sigma_L^2$ is the variance of the loss (which has to be guessed or estimated from training) and $d$ is half-width of a 95% confidence interval.

- Could report test the error, but then deploy whatever you train on the whole data. (Probably won’t be worse.)
Example - linear model

Training Data

## [1] "Estimated variance of errors: 0.168326238718343"
## [1] "Sample required for CI width of 0.2 (+- 0.1): 65"

Example - linear model

Testing Data

## TestMSE VarOfErrors StdOfSquaredErrors n StandardError CI_left CI_right
## 1 0.2261605 0.0980595 0.3131445 65 0.0388408 0.1500326
Choosing Performance Measures for Regression: Mean Errors

\[
\text{MSE} = n^{-1} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2
\]

\[
\text{RMSE} = \sqrt{n^{-1} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}
\]

\[
\text{MAE} = n^{-1} \sum_{i=1}^{n} |\hat{y}_i - y_i|
\]

I find MAE easier to interpret. (How far am I from the correct value, on average?) RMSE is at least in the same units as the \( y \).

Choosing Performance Measures for Regression: Mean Relative Error

\[
\text{MRE} = n^{-1} \sum_{i=1}^{n} \frac{|\hat{y}_i - y_i|}{|y_i|}
\]

Scales error according to magnitude of true \( y \). E.g., if MRE=0.2, then regression is wrong by 20\% of the value of \( y \), on average.

If this is appropriate for your problem then linear regression, which assumes additive error, may not be appropriate. Options include using a different model or regression on \( \log y \) rather than on \( y \).

https://en.wikipedia.org/wiki/Approximation_error#Formal_Definition

Extra slides - The Bootstrap

The Bootstrap

(AoS, p.110)

- CLT gives theoretical approximate sampling distribution of \( \bar{X}_n \).
- We could also estimate the sampling distribution of \( \bar{X}_n \) by drawing many datasets of size \( n \), computing \( \bar{X}_n \) on each, constructing histogram.
- This is impossible. \textit{But} we can use the data we have as a surrogate.

The Bootstrap

- Call our dataset \( D \).
- Draw \( B \) new datasets by sampling observations \textit{with replacement} from \( D \). (\( B \) is often at least 1000)
- Compute \( \bar{X}_n^{(b)} \) for each of the datasets.
- Use the histogram/empirical distribution of these “pretend” \( \bar{X} \) to determine confidence limits.
Bootstrap example

```r
library(boot)

bootstraps <- boot(faithful$eruptions,function(d,i){mean(d[i])},R=5000)

bootdata = data.frame(xbars=bootstraps$t);
limits = quantile(bootdata$xbars,c(0.025,0.975))

ggplot(bootdata, aes(x=xbars)) + 
  labs(y="Prop.") + geom_histogram(aes(y = ..density..)) +
  geom_errorbarh(aes(xmin=limits[[1]], xmax=limits[[2]], y=0),height=0.25,colour="red",size=2)
```
Reality Check

![Graph showing distributions of eruptions](image)