Performance Evaluation of Supervised Learning

Dan Lizotte

2018-09-25

Supervised Learning Framework (JWHT 2, HTF 2)

Training set: a set of labeled examples of the form

\[(x_1, x_2, \ldots, x_p, y),\]

where \(x_j\) are feature values and \(y\) is the output

- Task: Given a new \(x_1, x_2, \ldots, x_p\), predict \(y\)

What to learn: A function \(h : \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_p \rightarrow \mathcal{Y}\), which maps the features into the output domain

- Goal: Make accurate future predictions (on unseen data)

- From Reintroduction to Statistics, we saw how this goal is formalized in terms of Generalization Error

Types of Supervised Learning

- Problems are categorized by the type of output domain
  - If \(\mathcal{Y} = \mathbb{R}\), the problem is called regression
  - If \(\mathcal{Y}\) is a finite discrete set, the problem is called classification
  - If \(\mathcal{Y}\) has 2 elements, the problem is called binary classification

Supervised learning problem

- Given a data set \(D \subset (\mathcal{X} \times \mathcal{Y})^n\), find a function:

\[h : \mathcal{X} \rightarrow \mathcal{Y}\]

such that \(h(x)\) is a “good predictor” for the value of \(y\).

- \(h\) is called a predictive model or hypothesis

- Assumption: Dataset \(D\) is drawn from the same distribution that we will use to evaluate generalization error

Solving a supervised learning problem: optimisation-based approach

1. Decide what the input-output pairs are.
2. Decide how to encode inputs and outputs.
   
   This defines the input space \(\mathcal{X}\), and the output space \(\mathcal{Y}\).
3. Choose a class of models/hypotheses \(\mathcal{H}\).
4. Choose an error function (cost function) to define the best model in the class according to the training data.

5. Choose an algorithm for searching through the space of models to find the best one.

This approach is taken by many techniques, from Ordinary Least Squares (OLS) regression to Deep Learning.

Evaluating Performance
Evaluating Performance

Why do we need to evaluate performance?

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

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

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

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

Performance of a Fixed Hypothesis

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

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

• Given a set of data points \((x_i, y_i)\) not used to choose \(h\) that are realizations of \((X, Y)\), we can compute the test error

\[ \tilde{\ell}_{h,n} = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i) \]
Generalization error of hypotheses from last day

![Graph showing generalization error vs degree of polynomial. The x-axis represents the degree of polynomial, ranging from 1 to 9, and the y-axis represents the log of generalization error, ranging from 0 to 25. There are red dots indicating points on the graph, with a notable increase in error as the degree increases.]
Reminder: 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. Parameters are often expectations.

- **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 sampling distribution.

Bias

(AoS, p.90)

- The expected difference between estimator and estimand/parameter. For example,

\[ E[\bar{X}_n - \mu_X] \]

- Note: by convention \( \mu_X = E[X] \), the mean of r.v. \( X \).
- If 0, estimator is unbiased.
- Sometimes, \( \bar{x}_n > \mu_X \), sometimes \( \bar{x}_n < \mu_X \), but the long run average of these differences will be zero.

Variance

- The expected squared difference between estimator and its mean. For example,

\[ E[(\bar{X}_n - E[\bar{X}_n])^2] \]

- Positive for all non-trivial estimators. Higher variance means distribution of estimates are more “spread out.”
- Because \( \bar{X}_n \) is unbiased, we can write
\[ E[(\bar{X}_n - \mu_X)^2] \]

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

**Scenario - Cubic Polynomial, \( n = 10 \)**
Scenario - Cubic Polynomial, n = 100
Why is training error biased?

<table>
<thead>
<tr>
<th>Subset.Size</th>
<th>Error.on.16.Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1000</td>
</tr>
<tr>
<td>2</td>
<td>900</td>
</tr>
<tr>
<td>3</td>
<td>800</td>
</tr>
<tr>
<td>4</td>
<td>700</td>
</tr>
<tr>
<td>5</td>
<td>600</td>
</tr>
<tr>
<td>6</td>
<td>500</td>
</tr>
<tr>
<td>7</td>
<td>250</td>
</tr>
<tr>
<td>8</td>
<td>125</td>
</tr>
<tr>
<td>9</td>
<td>63</td>
</tr>
<tr>
<td>10</td>
<td>32</td>
</tr>
<tr>
<td>11</td>
<td>16</td>
</tr>
</tbody>
</table>

Implications for Performance Evaluation

- Training error is biased downward (optimistic on average)
  - For a fixed model space, the more data, the less bias we have
- Test error is always unbiased
- Variance of both training error and test error decreases as dataset size increases
Assessing the quality of an estimate

- By using knowledge of the bias and variance of estimators, we can understand how certain we are about estimates of performance
- Using statistics, we can
  - Determine how accurate our test-error-based estimates are
  - Make intelligent decisions about how much data we need for testing

Normal (Gaussian) Distribution

(AoS, p.28)

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

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

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 \( \bar{X}_n \), if we know \( \mu_{\bar{X}_n} \) and \( \sigma^2_{\bar{X}_n} \), then we can say a lot about how good it is.

Central Limit Theorem

(AoS, p.77)

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

\[
F_{\bar{X}_n}(\bar{x}) \approx \int_{-\infty}^{\bar{x}} \frac{1}{\sigma_{\bar{X}_n} \sqrt{2\pi}} e^{-\frac{(x-\mu_{\bar{X}_n})^2}{2\sigma_{\bar{X}_n}^2}}
\]

where

\[
\sigma_{\bar{X}_n} = \frac{\sigma}{\sqrt{n}}
\]

is called the standard error of \( \bar{X}_n \) 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 \( \bar{x}_{272} = 3.4877831 \).
- What is the probability of observing an \( \bar{x}_{272} \) that is within 10 seconds of the true mean?
Who cares?

By the C.L.T.,

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

\[ = 0.986 \]

Note! I estimated \( \sigma_X \) here. (Look up “t-test” for details.)

Confidence Intervals

(AoS, p.92)

- Typically, we specify confidence given by \( 1 - \alpha \)
- Use the sampling distribution to get
  
  \textit{an interval that traps the parameter (estimand) with probability} \( 1 - \alpha \).
- 95% C.I. for eruption mean is \((3.35, 3.62)\)
What a Confidence Interval Means
Effect of $n$ on width

Test Set Sample Sizes

- 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.

Rule-of-thumb size:

$$n = (1.96)^2 \frac{\sigma^2_L}{d^2}$$

where $\sigma^2_L$ is the variance of the losses (which has to be guessed or estimated from training) and $d$ is half-width of a 95% confidence interval.

- For linear regression, a rough estimate for $\sigma^2_L$ could be the variance of the squared errors in the training set. (But this will be biased downward as well.)
Example - linear model

### 

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

Example - linear model

### 

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

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

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

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

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

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

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

**Reporting guidelines**

- If possible, allocate some of your data for testing before you begin
- Use the sample size formula above to get a rough estimate of what you need
- Report the test error and confidence interval in your final results
- We will discuss other methods for when there isn’t enough data to allocate a test set this way