Hypothesis Testing

- Empirically evaluating accuracy of hypotheses: important activity in ML.

- Three questions:
  - Given observed accuracy over a sample set, how well does this estimate apply over additional samples?
  - Given a hypothesis outperforming another, how probable is it that this hypothesis is more accurate in general?
  - With limited data, how to learn and also estimate its accuracy?

- Use of statistical methods to put a bound on the error between the estimated and the true accuracy.

Evaluation of Performance of Learned $h$

- Want to decide whether to use $h$ or not: Want to understand the accuracy of the hypothesis learned from a limited-size training set.

- Evaluation may be part of the ML algorithm itself.

Issues

Learn hypothesis on limited data, and estimate future accuracy:

- Bias in the estimate:
  - The training data is a subset of the instance space, and may introduce bias: the estimated error may be different from the true error.

- Variance in the estimate:
  - Even though the estimate may be unbiased, there can be a large variance in the accuracy over different test sets.
  - Usually, smaller training sets lead to larger variance.

Trade-off Between Bias and Variance

- Less parameters $\rightarrow$ less accurate, but variance over different test sets is reduced.

- More parameters $\rightarrow$ more accurate, but variance over different test sets is increased.
**Topics**

- Evaluating hypotheses (estimate accuracy of a hypothesis).
- Compare accuracy of two hypotheses.
- Compare accuracy of two algorithms when data set is limited.

**Estimating Hypothesis Accuracy**

General setup:

- $X$: instance space.
- $D$: prob. distribution of encountering $x \in X$.

Task:

- Given hypothesis $h$ and data set of size $n$ from distribution $D$, what is the best estimate of the accuracy of $h$ on future instances from the same distribution?
- What is the probable error in the accuracy estimate?

**Probability Distribution of Sample Mean**

From instance space $X$, draw a small sample set $S_i$ of size $n$.

- For different sample sets $S_i$, the mean will differ:
  $$\mu_i \equiv \frac{1}{n} \sum_{x \in S_i} x$$

- The questions are:
  - Is $\mu_i = \mu_X$ (where $\mu_X$ is the true mean over $X$)?
  - How is $\mu_i$ distributed ($P(\mu)$, for $\mu \in \{\mu_1, \mu_2, \ldots \mu_n\}$)?

**Example of Sampling Distribution of the Mean**

$X = \{1, 2, 3, 4\}$, and each numbers are equally likely to occur (i.e., $D$ is a uniform distribution). Let’s sample with $n = 2$.

<table>
<thead>
<tr>
<th>Samples of size 2</th>
<th>Sample means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observation</td>
<td>1</td>
</tr>
<tr>
<td>1st \ 2nd</td>
<td>1,1</td>
</tr>
<tr>
<td>2</td>
<td>2,1</td>
</tr>
<tr>
<td>3</td>
<td>3,1</td>
</tr>
<tr>
<td>4</td>
<td>4,1</td>
</tr>
</tbody>
</table>

**Note:**

- From Kachigan (1991)

---

**References**


---
Sample Distribution vs. Sampling Distribution of the Mean

- Depending on how you sample your data, your sample mean can end up being different values.
- The sample mean has a distribution of its own centered at the actual population mean ($\sum_{x=\{1,2,3,4\}} \frac{1}{4}x = 2.5$).

True mean $\mu$ and sample mean $\mu_s$

- With a particular probability $p$, $\mu_s$ is within a particular range $r$ from the true mean $\mu$.
- In other words, if you pick any sample mean $\mu_s$, with the probability $p$, the true mean is within the range $r$.
- Given a fixed probability $p = 0.95$, the range $r$ is determined by the variance $\sigma_{\mu_s}$.

Sampling Distribution of the Mean

- Underlying distribution with mean $\mu$ and std $\sigma$.
- Distribution of sample mean $\mu_s$ has mean $\mu_{\mu_s} = \mu$ and std: $\sigma_{\mu_s} = \frac{\sigma}{\sqrt{n}}$, and tends to the normal distribution as $n$ grows.
- Interpretation:
  - When you get a particular sample mean $\mu_s$, you know it is distributed like $\sim \mathcal{N}(\mu, \sigma_{\mu_s})$.
  - With more samples, $\sigma_{\mu_s}$ reduces, so you’re more confident about your particular $\mu_s$ being close to the true mean $\mu$.

Sample Error and True Error

Sample error:
- Sample error of hypothesis $h$ based on sample set $S$ of size $n$:
  $$error_S(h) \equiv \frac{1}{n} \sum_{x \in S} \delta(f(x), h(x)),$$
  where $f(\cdot)$ is the target function, and $\delta(a, b) = 1$ if $a = b$ and $0$ if $a \neq b$.
- In other words, $error_S(h)$ is the mean error of hypothesis $h$.

True error:
- True error of hypothesis $h$ is the probability that $h$ will misclassify a single example drawn from the distribution $\mathcal{D}$:
  $$error_\mathcal{D}(h) \equiv \Pr_{x \in \mathcal{D}}[f(x) \neq h(x)]$$
Confidence Interval

- How good an estimator of $\text{error}_D(h)$ is provided by $\text{error}_S(h)$?
- Want to estimate true error based on sample $S$ of $n$ examples according to distribution $D$.
- $h$ commits $r$ errors: $\text{error}_S(h) = r/n$.
- With approx. 95% probability, true error is within the interval:
  $$\text{error}_S(h) \pm 1.96 \sqrt{\frac{\text{error}_S(h)(1 - \text{error}_S(h))}{n}}.$$

Confidence Interval Example

- $S$ of size $n = 40$.
- $h$ committing $r = 12$ errors.
- $\text{error}_S(h) = 12/40 = 0.30$ (mean error, or error rate).
- 95% confidence interval:
  $$0.30 \pm 1.96 \sqrt{\frac{0.3 \times (1.0 - 0.3)}{40}} = 0.30 \pm 0.14$$

Note: if $n$ is high, even when $r/n$ may be the same, the interval size would reduce.

Sampling Theory Basics: Summary

- Random variable: variable that can take on values with certain probability.
- Probability distribution: $\Pr(Y = y_i)$.
- Expected value: $E[Y] = \sum_i y_i \Pr(Y = y_i)$.
- Standard deviation: $\sqrt{\text{Var}(Y)}$.
- Binomial distribution: binary outcome, with probability $p$ of 0 and $(1 - p)$ for 1; Probability of $r$ 1’s with $n$ samples.
- Normal distribution
- Central limit theorem: sum of iid random variables tend to the normal distribution.
- Estimator is a random variable $Y$ that estimates parameter $p$.
- Estimation bias: $E(Y) - p$.
- $N\%$ confidence interval estimate of $p$: interval that includes true $p$ with $N\%$ probability.
Binomial Distribution: e.g., Coin Toss

- Outcome itself is described by a random variable $Y \in \{\text{Head}, \text{Tail}\}$.
- $P(Y = \text{Head}) = p$ and $P(Y = \text{Tail}) = (1 - p)$.
- Probability of observing $r$ heads out of $n$ coin tosses (this value corresponds to a random variable $R$):
  $$Pr(R = r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}.$$  
- $Pr(R = r)$ can be seen as the probability of observing $r$ errors in a sample size of $n$ (for binary target categories).

Mean and Variance in Binomical Distributions

- $E[Y] \equiv \sum_{i=1}^{n} y_i Pr(Y = y_i) = np$
- $Var[Y] \equiv E[(Y - E[Y])^2] = np(1-p)$

Errors, in Terms of Binominal Distribution

- $error_S(h) = \frac{r}{n}$
- $error_D = p$

Estimation Bias

- Estimation bias of an estimator $Y$ for a parameter $p$ is:
  $$E[Y] - p$$

Variance in Estimation

- $error_S(h) = \frac{r}{n}$
- Std[$r$] = $\sqrt{np(1-p)}$
- Std[$error_S(h)$] = $Std\left[\frac{r}{n}\right] = \frac{Std[r]}{n}$
  $$= \frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}}$$
  $$\approx \frac{\sqrt{error_S(h)(1-error_S(h))}}{n}$$

Normal Distribution

- Mean $E[X] = \mu$, and variance $Var[X] = \sigma^2$.
- Probability density:
  $$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2}$$
- Probability of falling between interval $[a, b]$:
  $$\int_{a}^{b} p(x)dx$$
- Central limit theorem: sum of a large number of iid random variables (the sum itself is a random variable) tends to Normal.
Confidence Interval in Normal Distributions

- $N\%$ of probability mass in Normal distributions are within:
  $$\mu \pm z_N \sigma.$$ 
- That means, a randomly drawn value $y$ will be within the above interval with a $N\%$ chance.
- In other words, if you pick any value $y$, with $N\%$ chance, the mean will be within the interval:
  $$y \pm z_N \sigma.$$ 

Calculating Confidence Intervals

1. Pick parameter $p$ to estimate
   - $\text{error}_D(h)$
2. Choose an estimator
   - $\text{error}_S(h)$
3. Determine probability distribution that governs estimator
   - Distribution of $\text{error}_S(h)$ can be approximated by Normal distribution when $n$ is large
4. Find interval $(L, U)$ such that $N\%$ of probability mass falls in the interval
   - Use table of $z_N$ values

Two-Sided vs. One-Sided Bounds

- Two-sided: Lower and upper bound with $100(1 - \alpha/2)\%$ confidence
- One-sided: Lower bound only (or upper bound only) with $100(1 - \alpha)\%$.
  - What is the probability that $\text{error}_D(h)$ is at most $U$?
Difference in Error of Two Hypotheses

Test $h_1$ on sample $S_1$, test $h_2$ on $S_2$

1. Pick parameter to estimate

$$d ≡ \text{error}_D(h_1) - \text{error}_D(h_2)$$

2. Choose an estimator

$$\hat{d} ≡ \text{error}_{S_1}(h_1) - \text{error}_{S_2}(h_2)$$

3. Determine probability distribution that governs estimator

$$\sigma_d ≈ \sqrt{\frac{\text{error}_{S_1}(h_1)(1 - \text{error}_{S_1}(h_1))}{n_1} + \frac{\text{error}_{S_2}(h_2)(1 - \text{error}_{S_2}(h_2))}{n_2}}$$

4. Find interval $(L, U)$ such that N% of probability mass falls in the interval

$$\hat{d} \pm zN\left(\frac{\text{error}_{S_1}(h_1)(1 - \text{error}_{S_1}(h_1))}{n_1} + \frac{\text{error}_{S_2}(h_2)(1 - \text{error}_{S_2}(h_2))}{n_2}\right)$$

Paired $t$-Test for Comparing $h_A$ and $h_B$

1. Partition data into $k$ disjoint test sets $T_1, T_2, \ldots, T_k$ of equal size, where this size is at least 30.

2. For $i$ from 1 to $k$, do

$$\delta_i \leftarrow \text{error}_{T_i}(h_A) - \text{error}_{T_i}(h_B)$$

3. Return the value $\bar{\delta}$, where

$$\bar{\delta} = \frac{1}{k} \sum_{i=1}^{k} \delta_i$$

$N\%$ confidence interval estimate for $d$:

$$\bar{\delta} \pm t_{N,(k-1)} \frac{s_{\bar{\delta}}}{\sqrt{k}}$$

$$s_{\bar{\delta}} = \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (\delta_i - \bar{\delta})^2}$$

Note: $\delta_i$ approximately Normally distributed, and $t$ differ for different sample size, as well as %.

Hypothesis Testing

- What is the prob. that $\text{error}_D(h_1) > \text{error}_D(h_2)$?

- Even if $\text{error}_{S_1}(h_1) > \text{error}_{S_2}(h_2)$, there is a chance that $\text{error}_D(h_1) < \text{error}_D(h_2)$.

- E.g., what is the chance of $d > 0$ when $\hat{d} = 0.1$ ($\text{error}_{S_1}(h_1) = 0.3$ and $\text{error}_{S_2}(h_2) = 0.2$)?
  - $\hat{d} < d + 0.1 = E[\hat{d}] + 0.1 = \mu_{\hat{d}} + 0.1$
  - $\hat{d} < \mu_d + 1.64 \times \sigma_{\hat{d}} = \mu_d + 1.64 \times 0.061$
  - $z_{90\%} = 1.64$ for two-sided interval, so the chance is 95%.

- Better to think how to reject the null hypothesis:
  - Null hypothesis $H_0: d = 0$
  - Alternative hypothesis $H_1: d > 0$ (must ensure $P(d < 0) = 0$)

Comparing learning algorithms $L_A$ and $L_B$

What we’d like to estimate:

$$E_{S \subset D}[\text{error}_D(L_A(S)) - \text{error}_D(L_B(S))]$$

where $L(S)$ is the hypothesis output by learner $L$ using training set $S$, i.e., the expected difference in true error between hypotheses output by learners $L_A$ and $L_B$, when trained using randomly selected training sets $S$ drawn according to distribution $D$.

But, given limited data $D_0$, what is a good estimator?

- could partition $D_0$ into training set $S$ and training set $T_0$, and measure

$$\text{error}_{T_0}(L_A(S_0)) - \text{error}_{T_0}(L_B(S_0))$$

- even better, repeat this many times and average the results (next slide)
Comparing learning algorithms $L_A$ and $L_B$

1. Partition data $D_0$ into $k$ disjoint test sets $T_1, T_2, \ldots, T_k$ of equal size, where this size is at least 30.

2. For $i$ from 1 to $k$, do
   - use $T_i$ for the test set, and the remaining data for training set $S_i$
     - $S_i \leftarrow \{D_0 - T_i\}$
     - $h_A \leftarrow L_A(S_i)$
     - $h_B \leftarrow L_B(S_i)$
     - $\delta_i \leftarrow error_{T_i}(h_A) - error_{T_i}(h_B)$

3. Return the value $\bar{\delta}$, where
   \[
   \bar{\delta} \equiv \frac{1}{k} \sum_{i=1}^{k} \delta_i
   \]