CS145: Lecture 20 Outline

- Bayesian Hypothesis Tests
- Non-Binary Hypothesis Tests
- Frequentist Hypothesis Tests
Bayesian Hypothesis Testing

Also known as classification, categorization, or discrimination.

We want to choose between two mutually exclusive hypotheses:
- $H=0$: Null hypothesis
- $H=1$: Alternative hypothesis

There is some prior probability of each hypothesis:
- Probability of $H=0$: $p_H(0) = q$
- Probability of $H=1$: $p_H(1) = 1 - q$

Observed data $X$ has a likelihood function under each hypothesis:
- Discrete data: $p_{X|H}(x | 0), p_{X|H}(x | 1)$
- Continuous data: $f_{X|H}(x | 0), f_{X|H}(x | 1)$

Formulas on following slides assume discrete $X$ for simplicity.
Bayesian hypothesis testing procedures assume that:

- The true value of the hypothesis is a random variable
- The prior distribution encodes previously observed data.

If no prior knowledge, set $p_H(0) = p_H(1) = 0.5$

We have a single new observation $X=x$, with likelihood $p_{X|H}(x | 0), p_{X|H}(x | 1)$

Compute posterior probability of hypothesis via Bayes rule:

$$p_{H|X}(h | x) = \frac{p_{X|H}(x | h)p_H(h)}{p_X(x)}$$

$$p_{H|X}(0 | x) + p_{H|X}(1 | x) = 1$$

$$p_X(x) = p_H(0)p_{X|H}(x | 0) + p_H(1)p_{X|H}(x | 1)$$

Typically both hypotheses have positive probability. How should we choose?
Loss Functions

We need to formalize the notion of the *cost of a mistake*:

\[ L(h, g) = \text{cost of predicting hypothesis } g \text{ when } h \text{ is true}. \]

Properties of standard *loss functions* used for hypothesis testing:

- Assume there is no loss for correct decisions:
  \[ L(0, 0) = L(1, 1) = 0 \]

- **Type I Error**: Positive loss for *false positives* or “false alarms”
  \[ L(0, 1) = \lambda_{01} > 0 \]

- **Type II Error**: Positive loss for *false negatives* or “missed detections”
  \[ L(1, 0) = \lambda_{10} > 0 \]

- Can encode “utilities” or “rewards” as negative losses
**Example: Biometric Identification**

\[ f_{X|H}(x \mid h) = \text{Features from phone’s camera, fingerprint sensor, …} \]

<table>
<thead>
<tr>
<th>Decision</th>
<th>( h=0: \text{ Authorized unlock} )</th>
<th>( h=1: \text{ Attacker} )</th>
</tr>
</thead>
</table>
| \( g = 0 \) | \( L(0, 0) = 0 \) | \( L(1, 0) = \lambda_{10} > 0 \)
  | \( \text{False negative:} \) \n  | \( \text{Attacker gains unauthorized access to phone!} \) |
| \( g = 1 \) | \( L(0, 1) = \lambda_{01} > 0 \) | \( L(1, 1) = 0 \)
  | \( \text{False positive:} \) \n  | \( \text{Enter biometric data again or enter passcode.} \) |
Bayesian Decision Theory

We are given both a \textit{probabilistic model} and a \textit{loss function}:

\textbf{Posterior distribution:}

\[ p_{H|X}(h \mid x) = \frac{p_{X|H}(x \mid h)p_H(h)}{p_X(x)} \]

\textbf{Loss function:}

\[ L(0, 1) = \lambda_{01} > 0 \quad \text{and} \quad L(1, 0) = \lambda_{10} > 0 \]

The optimal decision then \textit{minimizes the posterior expected loss:}

\[ \delta(x) = \arg \min_g E[L(h, g) \mid X = x] = \arg \min_g \sum_{h=0}^{1} L(h, g)p_{H|X}(h \mid x) \]
**Likelihood Ratio Tests**

Expected loss of guessing hypothesis $h=1$:

$$L(0, 1)p_{H|X}(0 \mid x) + L(1, 1)p_{H|X}(1 \mid x) = \lambda_{01}p_{H|X}(0 \mid x)$$

Expected loss of guessing hypothesis $h=0$:

$$L(0, 0)p_{H|X}(0 \mid x) + L(1, 0)p_{H|X}(1 \mid x) = \lambda_{10}p_{H|X}(1 \mid x)$$

The optimal decision then **minimizes the posterior expected loss**:

$$\delta(x) = \arg\min_g E[L(h, g) \mid X = x] = \arg\min_g \sum_{h=0}^1 L(h, g)p_{H|X}(h \mid x)$$
Likelihood Ratio Tests

Expected loss of guessing hypothesis $h=1$:

$$L(0, 1)p_{H|X}(0 \mid x) + L(1, 1)p_{H|X}(1 \mid x) = \lambda_{01}p_{H|X}(0 \mid x)$$

Expected loss of guessing hypothesis $h=0$:

$$L(0, 0)p_{H|X}(0 \mid x) + L(1, 0)p_{H|X}(1 \mid x) = \lambda_{10}p_{H|X}(1 \mid x)$$

It is optimal to decide $h=1$ if and only if:

$$\lambda_{01}p_{H|X}(0 \mid x) \leq \lambda_{10}p_{H|X}(1 \mid x)$$

$$\frac{p_{X|H}(x \mid 1)}{p_{X|H}(x \mid 0)} \geq \left( \frac{q}{1 - q} \right) \cdot \left( \frac{\lambda_{01}}{\lambda_{10}} \right)$$

$p_H(0) = q$
Minimizing Probability of Error

The general likelihood ratio test picks $h=1$ if and only if:

$$\lambda_{10} p_{H|X}(1 | x) \geq \lambda_{01} p_{H|X}(0 | x)$$

$$\frac{p_{X|H}(x | 1)}{p_{X|H}(x | 0)} \geq \left( \frac{q}{1 - q} \right) \cdot \left( \frac{\lambda_{01}}{\lambda_{10}} \right)$$

$p_{H}(0) = q$

If all errors are equally costly this simplifies: $\lambda_{10} = \lambda_{01} = 1$

$$p_{H|X}(1 | x) \geq p_{H|X}(0 | x)$$

$$\frac{p_{X|H}(x | 1)}{p_{X|H}(x | 0)} \geq \left( \frac{q}{1 - q} \right)$$

Pick hypothesis with larger posterior probability to minimize number of errors
Minimizing Probability of Error

The general likelihood ratio test picks $h=1$ if and only if:

$$\lambda_{10} p_{H|X}(1 \mid x) \geq \lambda_{01} p_{H|X}(0 \mid x)$$

$$\frac{p_{X|H}(x \mid 1)}{p_{X|H}(x \mid 0)} \geq \left(\frac{q}{1-q}\right) \cdot \left(\frac{\lambda_{01}}{\lambda_{10}}\right)$$

$p_H(0) = q$

If all errors are equally costly, and hypotheses have equal prior probability:

$$\lambda_{10} = \lambda_{01} = 1$$

$q = 0.5$

Pick hypothesis with larger likelihood to minimize number of errors
Example: Gaussian Hypothesis Tests

\[ p_H(0) = q \]

\[ f_{X \mid H}(x \mid 1) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2} \left( \frac{x - \mu_1}{\sigma_1} \right)^2} \]

\[ f_{X \mid H}(x \mid 0) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2} \left( \frac{x - \mu_0}{\sigma_0} \right)^2} \]

Assuming all errors are equally costly, we choose \( h=1 \) if:

\[ \frac{f_{X \mid H}(x \mid 1)}{f_{X \mid H}(x \mid 0)} \geq \left( \frac{q}{1 - q} \right) \]

\[ c = \log \left( \frac{q}{1 - q} \right) \]

\[ \log(f_{X \mid H}(x \mid 1)) - \log(f_{X \mid H}(x \mid 0)) \geq c \]
Example: Gaussian Hypothesis Tests

\[
\log(f_{X|H}(x | i)) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma_i^2) - \frac{1}{2} \left(\frac{x - \mu_i}{\sigma_i}\right)^2
\]

Suppose that \(\sigma_1 = \sigma_0 = \sigma\) and \(\mu_1 > \mu_0\):

\[
\log(f_{X|H}(x | 1)) - \log(f_{X|H}(x | 0)) \geq c
\]

\[
-\frac{1}{2\sigma^2} (x - \mu_1)^2 + \frac{1}{2\sigma^2} (x - \mu_0)^2 \geq c
\]

With some algebra, we choose \(h=1\) if:

\[
x \geq \frac{\mu_1 + \mu_0}{2} + \frac{\sigma^2 c}{\mu_1 - \mu_0}
\]
Example: Gaussian Hypothesis Tests

\[
\log(f_{X|H}(x \mid i)) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma_i^2) - \frac{1}{2} \left(\frac{x - \mu_i}{\sigma_i}\right)^2
\]

Suppose that \(\mu_1 = \mu_0 = 0\) and \(\sigma_1 > \sigma_0\):

\[
\log(f_{X|H}(x \mid 1)) - \log(f_{X|H}(x \mid 0)) \geq c
\]

\[
-\frac{x^2}{2\sigma_1^2} + \frac{x^2}{2\sigma_0^2} - \frac{1}{2} \log(\sigma_1^2) + \frac{1}{2} \log(\sigma_0^2) \geq c
\]

With some algebra, we choose \(h=1\) if:

\[
x^2 \geq \frac{2\sigma_1^2\sigma_0^2}{\sigma_1^2 - \sigma_0^2} \left(c + \log\left(\frac{\sigma_1}{\sigma_0}\right)\right)
\]
\[
\log(f_{X|H}(x \mid i)) = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_i| - \frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)
\]

- Decision boundary is always a \textit{quadratic function}
- If classes have same covariance, decision boundary is a \textit{linear function}
Continuous Spam Detection Features

Detecting Spammers with SNARE: Spatio-temporal Network-level Automatic Reputation Engine

Hao et al., USENIX 2009
CS145: Lecture 20 Outline

- Bayesian Hypothesis Tests
- Non-Binary Hypothesis Tests
  *Advanced topic not covered in homeworks or exams.*
- Frequentist Hypothesis Tests
Non-Binary Hypothesis Tests

We want to choose between $M$ mutually exclusive hypotheses:

$\displaystyle L(h, g) = \text{cost of predicting hypothesis } g \text{ when } h \text{ is true.} \quad L(h, h) = 0$

There is some prior probability of each hypothesis: \( p_H(h) \)

Observed data $X$ has a likelihood function under each hypothesis:

- Discrete data: \( p_X|H(x \mid h), \quad h = 1, \ldots, M \)
- Continuous data: \( f_X|H(x \mid h), \quad h = 1, \ldots, M \)

The optimal decision then minimizes the posterior expected loss:

\[
\delta(x) = \arg \min_g E[L(h, g) \mid X = x] = \arg \min_g \sum_{h=1}^{M} L(h, g) p_{H|X}(h \mid x)
\]
Suppose that all errors are equally costly:
\[ L(h, g) = 1 \text{ if } h \neq g \]
\[ L(h, h) = 0 \]

The optimal decision then picks the most probable hypothesis:
\[ \delta(x) = \arg \max_g p_{H|X}(g \mid x) \]

If it is also true that all hypotheses are equally likely:
\[ \delta(x) = \arg \max_g p_{X|H}(x \mid g) \text{ in cases where } p_H(g) = \frac{1}{M} \]

*This is the maximum likelihood decision rule.*
Example: Multiple Gaussian Classes

\[
\log(f_{X|H}(x \mid i)) = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_i| - \frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)
\]
LS145: Lecture 20 Outline

- Bayesian Hypothesis Tests
- Non-Binary Hypothesis Tests
- Frequentist Hypothesis Tests
Frequentist Hypothesis Testing

Also known as classification, categorization, or discrimination.

We want to choose between two mutually exclusive hypotheses:

- \( h=0 \): Null hypothesis \( H_0 \)
- \( h=1 \): Alternative hypothesis \( H_1 \)

Observed data \( X \) has a likelihood function under each hypothesis:

- Discrete data: \( p_X(x; 0), p_X(x; 1) \)
- Continuous data: \( f_X(x; 0), f_X(x; 1) \)

Parameterized by \( h \), not conditioned on \( h \)!

Formulas on following slides assume discrete \( X \) for simplicity.

The frequentist approach has one major conceptual difference:

- The true hypothesis is unknown, but fixed and deterministic
- The hypothesis is not a random variable and there is no prior
**Decision Rule or Classification Rule:**

If \( x \in R \) guess \( g(x) = 1 \).

If \( x \in R^c \) guess \( g(x) = 0 \).

**Type I Error:** False positives or “false alarms”:

\[ g(x) = 1 \text{ but } h = 0. \]

**Type II Error:** False negatives or “missed detections”:

\[ g(x) = 0 \text{ but } h = 1. \]

- The frequentist approach thinks of the data as random
- If we run our decision-making experiment many times under the same conditions, what is the **frequency** of each error?
Frequentist Error Probabilities

**Decision Rule or Classification Rule:**

If \( x \in R \) guess \( g(x) = 1 \).
If \( x \in R^c \) guess \( g(x) = 0 \).

**Type I Error:** *False positives* or “false alarms”:
\( g(x) = 1 \) but \( h = 0 \).

**Type II Error:** *False negatives* or “missed detections”:
\( g(x) = 0 \) but \( h = 1 \).

- Frequentist probability of *false positives*:
\[
\alpha(R) = P(X \in R; h = 0) = \sum_{x \in R} p_X(x; 0)
\]

- Frequentist probability of *false negatives*:
\[
\beta(R) = P(X \notin R; h = 1) = 1 - \sum_{x \in R} p_X(x; 1)
\]
Consider the following \textit{degenerate decision rules}:

- Always predict $g(x)=1$, ignoring $x$
- Always predict $g(x)=0$, ignoring $x$
- Flip a coin and randomly pick a hypothesis, ignoring $x$

\textit{It’s easy to make one type of error small, if you don’t care about the other. Being always wrong is as hard as being always right! Why?}
Recall that the likelihood ratio is defined as:

\[
L(x) = \frac{p_X(x; 1)}{p_X(x; 0)} \quad \text{(discrete data)} \quad L(x) = \frac{f_X(x; 1)}{f_X(x; 0)} \quad \text{(continuous data)}
\]

A likelihood ratio test (LRT) then chooses \( g(x) = 1 \) if and only if

\[
L(x) > \xi
\]

Degree of freedom: critical value or threshold \( \xi \)
Classification via Likelihood Ratio Tests

Decision Rule or Classification Rule:

\[ g(x) = 1 \text{ if } L(x) > \xi, \]
\[ g(x) = 0 \text{ if } L(x) \leq \xi. \]

- Convention: Choose a **target false positive probability**
  \[ \alpha = P(g(X) = 1; h = 0) \]
- Tune the **LRT threshold to match** the target false positive rate:
  \[ P(L(X) > \xi; h = 0) = \alpha \]
- The **false negative probability** \( \beta \) can then be computed
### Example: Spam Classification

\[ p(x|h) = \]  

*Model of words in email: naïve Bayes, Markov chain, …*

<table>
<thead>
<tr>
<th>Decision</th>
<th>( h=0: ) Ham (not spam)</th>
<th>( h=1: ) Spam</th>
</tr>
</thead>
<tbody>
<tr>
<td>( g = 0 )</td>
<td>( 1 - \alpha )</td>
<td>( \beta )</td>
</tr>
<tr>
<td>\text{True negative rate:}</td>
<td>Correct decision!</td>
<td>\text{False negative rate:}</td>
</tr>
<tr>
<td>( g = 1 )</td>
<td>( \alpha )</td>
<td>( 1 - \beta )</td>
</tr>
<tr>
<td>\text{False positive rate:}</td>
<td>Fraction of real emails placed in spam folder.</td>
<td>\text{True positive rate:}</td>
</tr>
</tbody>
</table>
Example: Gaussian Hypothesis Tests

\[ f_X(x; h = i) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{1}{2} \left( \frac{x - \mu_i}{\sigma} \right)^2} \]

\[ \mu_0 = 0, \mu_1 = 1 \]

\[ L(x) = \frac{f_X(x; 1)}{f_X(x; 0)} = \exp \left\{ \frac{2x - 1}{2\sigma^2} \right\} \]

- For any choice of the critical value or threshold find the false positive rate:

\[ R = \{ x \mid L(x) > \xi \} = \{ x \mid x > \gamma \} \quad \gamma = \sigma^2 \log(\xi) + \frac{1}{2} \]

\[ \alpha = P(X > \gamma; h = 0) = P \left( \frac{X}{\sigma} > \frac{\gamma}{\sigma}; h = 0 \right) = 1 - \Phi \left( \frac{\gamma}{\sigma} \right) \]

- Tune via normal CDF and compute false negative rate:

\[ \beta = P(X \leq \gamma; h = 1) = P \left( \frac{X - 1}{\sigma} \leq \frac{\gamma - 1}{\sigma}; h = 1 \right) = \Phi \left( \frac{\gamma - 1}{\sigma} \right) \]
Consider any fixed likelihood ratio test with threshold $\xi$

$$P(L(X) > \xi; h = 0) = \alpha \quad P(L(X) \leq \xi; h = 1) = \beta$$

Then for any other possible classifier, with rejection region $R$,
if $P(X \in R; h = 0) \leq \alpha$ then $P(X \notin R; h = 1) \geq \beta$

For any target false positive rate, the LRT has the best possible false negative rate!
Bayesian & Frequentist Tests

Discrete or Continuous Likelihood Ratios:

\[ L(x) = \frac{p_X(x; 1)}{p_X(x; 0)} \quad L(x) = \frac{f_X(x; 1)}{f_X(x; 0)} \]

Decision Rule or Classification Rule:

\[ g(x) = 1 \text{ if } L(x) > \xi, \]
\[ g(x) = 0 \text{ if } L(x) \leq \xi. \]

**Bayesian:** Set threshold to minimize expected loss \( L(h, g) \)

\[ \xi = \left( \frac{p_H(0)}{p_H(1)} \right) \cdot \left( \frac{L(0, 1)}{L(1, 0)} \right) \]

**Frequentist:** Set threshold to control false positive rate

\[ P(L(X) > \xi; h = 0) = \alpha \]
Bayesian & Frequentist Tests

**Likelihood Ratio Tests are Optimal:** Neyman-Pearson Lemma

- Any likelihood ratio test is a special case of an optimal Bayesian decision rule, for appropriate prior probabilities
- By optimality of Bayesian decision rule, it is impossible for another classifier to have a lower average probability of error
- Allows proof by contradiction (see B&T Section 9.3)

\[
L(x) = \frac{p_X(x; 1)}{p_X(x; 0)} \\
L(x) = \frac{f_X(x; 1)}{f_X(x; 0)}
\]

**Decision Rule or Classification Rule:**

\[
g(x) = 1 \text{ if } L(x) > \xi, \\
g(x) = 0 \text{ if } L(x) \leq \xi.
\]
Bayesian & Frequentist Tests

Discrete or Continuous Likelihood Ratios:

\[ L(x) = \frac{p_X(x; 1)}{p_X(x; 0)} \quad \text{and} \quad L(x) = \frac{f_X(x; 1)}{f_X(x; 0)} \]

Decision Rule or Classification Rule:

\[ g(x) = 1 \text{ if } L(x) > \xi, \]
\[ g(x) = 0 \text{ if } L(x) \leq \xi. \]

➢ Bayesian:
  ➢ Fixed: The single data set we have observed.
  ➢ Random: The true value of the hypothesis, given our partial knowledge.

➢ Frequentist:
  ➢ Fixed: The true (but unknown) state of the hypothesis in the world.
  ➢ Random: The data, over many hypothetical repetitions of experiment.