CS145: Lecture 21 Outline

- Frequentist Hypothesis Tests
- Maximum Likelihood Parameter Estimation
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: \( f_X(x; 0) \), \( f_X(x; 1) \)

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
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”:

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

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

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

- **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:**

\[ \beta(R) = P(X \notin R; h = 1) = 1 - \sum_{x \in R} p_X(x; 1) \]
Recall that the likelihood ratio is defined as:

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

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

\[ L(x) > \xi \]

\[ R = \{ x \mid L(x) > \xi \} \]

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: 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:

\[ \beta = P(X \leq \gamma; h = 1) = P \left( \frac{X - 1}{\sigma} \leq \frac{\gamma - 1}{\sigma}; h = 1 \right) = \Phi \left( \frac{\gamma - 1}{\sigma} \right) \]
Comparison: Bayesian 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 \( \mu_0 = 0, \mu_1 = 1 \):

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

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

\[
c = \log \left( \frac{p_H(0)\lambda_{01}}{p_H(1)\lambda_{10}} \right)
\]

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

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

\[\mu_0 = 0 \quad \mu_1 = 1\]
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:**

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

- **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*

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

Maximum Likelihood Parameter Estimation
Statistical Inference Problems

**Hypothesis Testing:** How do I categorize “test data”?

- Two (or more) mutually exclusive hypotheses: $H=0$ or $H=1$?
- The distribution of the data under each hypothesis is known:
  
  
  \[ p_{X|H}(x | 0), \quad p_{X|H}(x | 1) \]
  
  \[ f_{X|H}(x | 0), \quad f_{X|H}(x | 1) \]

  - **Goal:** Choose between hypotheses

**Estimation:** How do I learn from “training data”?

- We have $n$ independent observations sampled from some unknown probability distribution: $x_1, x_2, \ldots, x_n$
- We assume the distribution of our data lives in some family, but don’t know the right parameter values $\theta$

  - **Goal:** Learn parameters that best “explain” the observations
Example: Bernoulli Distribution

- A *Bernoulli* or *indicator* random variable $X$ has one parameter:
  \[ p_X(1) = \theta, \quad p_X(0) = 1 - \theta, \quad \mathcal{X} = \{0, 1\} \]

- The *probability mass function* for an observation $x_i$ equals:
  \[ p_X(x_i; \theta) = \theta^{x_i} (1 - \theta)^{1-x_i} \]
  \[ \log p_X(x_i; \theta) = x_i \log(\theta) + (1 - x_i) \log(1 - \theta) \]

- **Goal:** Estimate $\theta$ from $n$ independent observations:
  \[ P(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} p_X(x_i; \theta) \]
Example: Uniform Distribution

- A continuous *uniform distribution* between 0 and $\theta$ has the following probability density function:

\[
   f_X(x_i; \theta) = \frac{1}{\theta} \quad \text{if } 0 \leq x_i \leq \theta,
\]

\[
   f_X(x_i; \theta) = 0 \quad \text{if } x_i < 0 \text{ or } x_i > \theta.
\]

- **Goal:** Estimate $\theta$ from $n$ independent observations:

\[
   f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_X(x_i; \theta)
\]
Example: Gaussian Distribution

A univariate \textit{Gaussian} distribution is parameterized by its mean and variance:

\[
f_X(x_i; \theta) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2}
\]

\[
\theta = \{ \mu, \sigma^2 \}
\]

Goal: Estimate \( \theta \) from \( n \) independent observations:

\[
f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_X(x_i; \theta)
\]
Maximum Likelihood (ML) Estimation

- Suppose I have \( n \) independent observations sampled from some unknown probability distribution: \( x = \{x_1, x_2, \ldots, x_n\} \)

- Suppose I have two candidate parameter estimates where:

\[
p_X(x; \theta_1) > p_X(x; \theta_2)
\]

Given no other information, choose the higher likelihood model!

- The maximum likelihood (ML) parameter estimate is defined as:

\[
\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} p_X(x_i; \theta) = \arg\max_{\theta} \sum_{i=1}^{n} \log p_X(x_i; \theta) \quad \text{Discrete Observations}
\]

\[
\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} f_X(x_i; \theta) = \arg\max_{\theta} \sum_{i=1}^{n} \log f_X(x_i; \theta) \quad \text{Continuous Observations}
\]
Finding ML Estimates

- For many practical models, the log-likelihood is a smooth and continuous function of the parameters:

\[ L(\theta) = \sum_{i=1}^{n} \log p_X(x_i; \theta) \]

Maximum will occur at a point where derivative equals zero!

- The maximum likelihood (ML) parameter estimate is defined as:

\[ \hat{\theta} = \arg \max_{\theta} \prod_{i=1}^{n} p_X(x_i; \theta) = \arg \max_{\theta} \sum_{i=1}^{n} \log p_X(x_i; \theta) \quad \text{Discrete Observations} \]

\[ \hat{\theta} = \arg \max_{\theta} \prod_{i=1}^{n} f_X(x_i; \theta) = \arg \max_{\theta} \sum_{i=1}^{n} \log f_X(x_i; \theta) \quad \text{Continuous Observations} \]
Example: Bernoulli Distribution

- A Bernoulli or indicator random variable $X$ has one parameter:

$$p_X(x_i; \theta) = \theta^{x_i} (1 - \theta)^{1-x_i} \quad x_i \in \{0, 1\}$$

- The maximum likelihood (ML) estimate maximizes:

$$L(\theta) = \sum_{i=1}^{n} x_i \log(\theta) + (1 - x_i) \log(1 - \theta)$$

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{ML: Empirical fraction of successes!}$$
Example: Exponential Distribution

- A \textit{geometric} random variable $X$ has parameter:
  \[
  f_X(x_i; \theta) = \theta e^{-\theta x_i}, \quad x_i \geq 0.
  \]

- The \textit{maximum likelihood (ML)} estimate maximizes:
  \[
  L(\theta) = \sum_{i=1}^{n} \log(\theta) - \theta x_i
  \]

  \[
  \hat{\theta} = \left( \frac{1}{n} \sum_{i=1}^{n} x_i \right)^{-1}
  \]

  \textit{ML: Match model to empirical mean!}
A univariate *Gaussian* distribution is:

\[
f_X(x_i; \theta) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2}
\]

The *ML* estimate maximizes:

\[
L(\mu, \sigma) = \sum_{i=1}^{n} -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{1}{2} \left( \frac{x_i - \hat{\mu}}{\sigma} \right)^2
\]

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2
\]
A continuous uniform distribution between 0 and \( \theta \):

\[
f_X(x_i; \theta) = \frac{1}{\theta} \quad \text{if} \ 0 \leq x_i \leq \theta,
\]

\[
f_X(x_i; \theta) = 0 \quad \text{if} \ x_i < 0 \text{ or } x_i > \theta.
\]

The maximum likelihood (ML) estimate maximizes:

\[
L(\theta) = \prod_{i=1}^{n} f_X(x_i; \theta)
\]

Cannot take logarithm because density can equal exactly zero.

Optimal to choose smallest \( \theta \) under which the data has positive probability:

\[
\hat{\theta} = \max\{x_1, x_2, \ldots, x_n\}
\]
Remainder: Convergence in Probability

Convergence in Probability
Let $Y_1, Y_2, \ldots$ be a sequence of random variables (not necessarily independent), and let $a$ be a real number. We say that the sequence $Y_n$ converges to $a$ in probability, if for every $\epsilon > 0$, we have

$$\lim_{n \to \infty} P(|Y_n - a| \geq \epsilon) = 0.$$ 

Weak Law of Large Numbers:

$X_1, X_2, \ldots$ i.i.d. finite mean $\mu$ and variance $\sigma^2$

$$M_n = \frac{X_1 + \cdots + X_n}{n} \quad E[M_n] = \mu$$

The “empirical mean” converges to true mean in probability:

$$\lim_{n \to \infty} P(|M_n - \mu| \geq \epsilon) = 0 \quad \text{for any } \epsilon > 0$$
Consistency of ML Estimators

\[ \hat{\theta}_n = \arg \max_{\theta} \prod_{i=1}^{n} f_X(x_i; \theta) = \arg \max_{\theta} \sum_{i=1}^{n} \log f_X(x_i; \theta) \]

- An estimator is **consistent** if the sequence of estimates \( \hat{\theta}_n \) converges to the true parameter in probability, for any true \( \theta \)
  \[ \lim_{n \to \infty} P( |\hat{\theta}_n - \theta| \geq \epsilon ) = 0 \quad \text{for any } \epsilon > 0 \]
- Under “mild conditions” that are true for most distributions, the **ML parameter estimates are always consistent**
  
  **Examples:** Gaussian distribution, uniform distribution, …

- The ML estimator also satisfies a **central limit theorem**, and for large \( n \) has provably small variance (“efficiency”)