Announcements

- Proposals graded
Hypothesis testing

Machine Learning – CSE546
Kevin Jamieson
University of Washington

October 30, 2018
You are Amazon and wish to detect transactions with stolen credit cards.

For each transaction we observe a **feature vector** $X$: 
{ email-address, age of account, anonymous PO box, price of items, copies of purchased item, etc. } and the transaction is either **real** ($Y=0$) or **fraudulent** ($Y=1$)

**Hypothesis testing:**

\[
\begin{align*}
H_0: & \quad X \sim P_0 \\
H_1: & \quad X \sim P_1 \\
Pk & = \mathbb{P}(X = x | Y = k)
\end{align*}
\]

Your job is to build a (possibly randomized) decision function $\delta(x) \in \{0, 1\}$
Anomaly detection

Hypothesis testing:

\[ H_0: \ X \sim P_0 \]
\[ H_1: \ X \sim P_1 \]

Your job is to build a (possibly randomized) decision function \( \delta(x) \in \{0, 1\} \)

Bayesian Hypothesis Testing:

Assume \( \mathbb{P}(Y = 1) = \pi \)
\[ \mathbb{P}(X = x) = \pi P_1(x) + (1 - \pi)P_0(x) \]

\[ P_k = \mathbb{P}(X = x | Y = k) \]

\[ \arg \min_{\delta} \mathbb{P}_{XY}(Y \neq \delta(X)) \]
Anomaly detection

Hypothesis testing:

$H_0$: $X \sim P_0$

$H_1$: $X \sim P_1$

$P_k = \mathbb{P}(X = x|Y = k)$

Your job is to build a (possibly randomized) decision function $\delta(x) \in \{0, 1\}$

Minimax Hypothesis Testing:

$\arg \min_\delta \max \{\mathbb{P}(\delta(X) = 0|Y = 1), \mathbb{P}(\delta(X) = 1|Y = 0)\}$
Anomaly detection

Hypothesis testing:

\( \text{H0: } X \sim P_0 \)
\( \text{H1: } X \sim P_1 \)

Your job is to build a (possibly randomized) decision function \( \delta(x) \in \{0, 1\} \)

Neyman-Pearson Hypothesis Testing:

\[ \arg \max_{\delta} \Pr(\delta(X) = 1|Y = 1), \text{ subject to } \Pr(\delta(X) = 1|Y = 0) \leq \alpha \]
Neyman-Pearson Testing

Hypothesis testing:

H0: $X \sim P_0$

H1: $X \sim P_1$

Neyman-Pearson Hypothesis Testing:

$P_k = \mathbb{P}(X = x | Y = k)$

\[
\arg\max_{\delta} \mathbb{P}(\delta(X) = 1 | Y = 1), \text{ subject to } \mathbb{P}(\delta(X) = 1 | Y = 0) \leq \alpha
\]

Theorem: The optimal test $\delta^*$ has the form

$\mathbb{P}(\delta^*(X) = 1) = \begin{cases} 1 & \text{if } \frac{P_1(x)}{P_0(x)} > \eta \\ \gamma & \text{if } \frac{P_1(x)}{P_0(x)} = \eta \\ 0 & \text{if } \frac{P_1(x)}{P_0(x)} < \eta \end{cases}$

and satisfies $\mathbb{P}(\delta^*(X) = 1 | Y = 0) = \alpha$
Neyman-Pearson Testing

Hypothesis testing:
\[ \text{H}_0: \ X \sim P_0 \quad \quad P_k = \mathbb{P}(X = x|Y = k) \]
\[ \text{H}_1: \ X \sim P_1 \]

Neyman-Pearson Hypothesis Testing:
\[ \arg \max_{\delta} \mathbb{P}(\delta(X) = 1|Y = 1), \quad \text{subject to} \quad \mathbb{P}(\delta(X) = 1|Y = 0) \leq \alpha \]

Example:
ROC Curve

Hypothesis testing:

H₀: \( X \sim P₀ \)

H₁: \( X \sim P₁ \)

\[ P_k = \mathbb{P}(X = x|Y = k) \]

Prob of False Alarm

\[ \mathbb{P}(\delta(X) = 1|Y = 0) \]

Prob of Detection

\[ \mathbb{P}(\delta(X) = 1|Y = 1) \]
p-values

Machine Learning – CSE546
Kevin Jamieson
University of Washington

October 30, 2018
You are Amazon and wish to detect transactions with stolen credit cards.

For each transaction we observe a feature vector $X$: { email-address, age of account, anonymous PO box, price of items, copies of purchased item, etc. } and the transaction is either real ($Y=0$) or fraudulent ($Y=1$)

**Hypothesis testing:**

- $H_0: X \sim P_0$
- $H_1: X \sim P_1$

$$P_k = \mathbb{P}(X = x \mid Y = k)$$

Your job is to build a (possibly randomized) decision function $\delta(x) \in \{0, 1\}$

Natural to have model for $P_0$ (regular purchases). But what if we have no model for $P_1$ since people are strategic?
Hypothesis testing:

$H_0: X \sim P_0$

$H_1: X \sim P_1$

$p$-value

$P_k = \mathbb{P}(X = x | Y = k)$

Your job is to build a (possibly randomized) decision function $\delta(x) \in \{0, 1\}$

**Definition**

$p$-value: probability of finding the observed, or more extreme, results when the null hypothesis $H_0$ is true (e.g., $X \sim P_0$)

**Definition**

$p$-value: a uniformly distributed random variable under the null hypothesis (e.g., $X \sim P_0$)

**WARNING**: A small $p$-value is **NOT** evidence that $H_1$ is true.
**p-value**

**Hypothesis testing:**

\[ H_0: \quad X \sim P_0 \]

\[ H_1: \quad X \sim P_1 \]

Your job is to build a (possibly randomized) decision function \( \delta(x) \in \{0, 1\} \)

**Definition** p-value: a uniformly distributed random variable under the null hypothesis (e.g., \( X \sim P_0 \))

\[
P_0(x) = \mathcal{N}(x; \mu_0, \sigma^2)
\]

**Observe:** \( x_i \in \mathbb{R} \)

**p-value:** \( p_i = P_0(X \geq x_i) \)

\[
= \int_{x=x_i}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_0)^2}{2\sigma^2}} \, dx
\]
p-value: used the **right** way

**Hypothesis testing:**

H0: $X \sim P_0$

H1: $X \sim P_1$

Your job is to build a (possibly randomized) decision function $\delta(x) \in \{0, 1\}$

**Definition** p-value: a uniformly distributed random variable under the null hypothesis (e.g., $X \sim P_0$)

**Set:** $\alpha = .05$

**Observe:** $x_i \in \mathbb{R}$

**p-value:** $p_i = P_0(X \geq x_i)$

**Test:** If $p_i \leq \alpha$ then reject the null hypothesis H0
p-value: used the **wrong** way

**Hypothesis testing:**

\[
\begin{align*}
H_0 &: X \sim P_0 \\
H_1 &: X \sim P_1 \\
\end{align*}
\]

\[
P_k = P(X = x | Y = k)
\]

Your job is to build a (possibly randomized) decision function \( \delta(x) \in \{0, 1\} \)

**Definition** p-value: a uniformly distributed random variable under the null hypothesis (e.g., \( X \sim P_0 \))

**Set:** \( \alpha = .05 \)

**Observe:** \( x_i \in \mathbb{R} \)

**p-value:** \( p_i = P_0(X \geq x_i) \)

**Test:** If \( p_i \leq \alpha \) then **reject** the null hypothesis \( H_0 \)

**BAD** If \( p_i > \alpha \) repeat the experiment with new \( x_i \) until \( p_i \leq \alpha \)
p-value: used the **wrong** way

Each day $i=1,2,...$ you measure an iid $x_i \sim \mathcal{N}(\mu, 1)$

$H_0: \mu = 0$

Under $H_0$ the statistic $Z_i = \frac{1}{\sqrt{i}} \sum_{j=1}^{i} x_j \sim \mathcal{N}(0, 1)$

\[
p_i = \frac{1}{2\pi} \int_{z=z_i}^{\infty} e^{-z^2/2} \, dz
\]
Multiple testing

Machine Learning – CSE546
Kevin Jamieson
University of Washington

October 30, 2018
Case study in adaptive sampling tradeoffs


Each gene $i=1,2,\ldots,n$ you measure an $x_i \sim \mathcal{N}(\mu_i, 1)$

$H_0(i): \mu_i = 0$

Consider procedure for individual hypothesis testing:

Set: $\alpha = .05$

Observe: $x_i \in \mathbb{R}$

p-value: $p_i = P_0(X \geq x_i)$

Test: If $p_i \leq \alpha$ then reject the null hypothesis $H_0$

Under $H_0$, how many genes do we expect to reject the null hypothesis?
Multiple Testing

If we make $n$ rejections individually at level $\alpha$

$$I_0 = \{i : H0(i) \text{ is true}\}$$

$$\mathbb{E}[\sum_{i \in I_0} 1\{p_i \leq \alpha\}] = \sum_{i \in I_0} \mathbb{P}(p_i \leq \alpha) = |I_0| \alpha$$

That’s a lot of false alarms!
Multiple Testing - FWER

Family-wise error rate $FWER = \mathbb{P}(\text{reject any true null})$

$$I_0 = \{i : H_0(i) \text{ is true}\}$$

Bonferroni rule: Reject $i$ if $p_i \leq \alpha/n$

$$FWER = \mathbb{P}\left(\bigcup_{I_0}\{p_i \leq \alpha/n\}\right) =$$
Multiple Testing - FDR

False discovery rate FDR = $\mathbb{E} \left[ \frac{|I_0 \cap R|}{|R|} \right]$

$I_0 = \{ i : H0(i) \text{ is true} \}$

Benjamini-Hochberg procedure:

Sort $p$-values such that $p(1) \leq p(2) \leq \cdots \leq p(n)$

$i_{\text{max}} = \max \{ i : p(i) \leq \frac{i}{n} \alpha \}$

$R = \{ i : i \leq i_{\text{max}} \}$

Theorem: BH($\alpha$) satisfies $\text{FDR} \leq \alpha$
Bayesian Methods

Machine Learning – CSE546
Kevin Jamieson
University of Washington

October 30, 2018
MLE Recap - coin flips

- **Data**: sequence \( D = (HHTHT...) \), \( k \) heads out of \( n \) flips
- **Hypothesis**: \( P(\text{Heads}) = \theta, \ P(\text{Tails}) = 1-\theta \)

\[
P(D|\theta) = \theta^k (1 - \theta)^{n-k}
\]

- Maximum likelihood estimation (MLE): Choose \( \theta \) that maximizes the probability of observed data:

\[
\hat{\theta}_{MLE} = \arg \max_{\theta} P(D|\theta)
= \arg \max_{\theta} \log P(D|\theta)
= \frac{k}{n}
\]
What about prior

- **Billionaire**: Wait, I know that the coin is “close” to 50-50. What can you do for me now?
- **You say**: I can learn it the Bayesian way…
Bayesian Learning

- Use Bayes rule:

\[ P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})} \]

- Or equivalently:

\[ P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta) \]
Bayesian Learning for Coins

\[ P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta)P(\theta) \]

- Likelihood function is simply Binomial:
  \[ P(\mathcal{D} | \theta) = \theta^\alpha_H (1 - \theta)^\alpha_T \]

- What about prior?
  - Represent expert knowledge

- Conjugate priors:
  - Closed-form representation of posterior
  - For Binomial, conjugate prior is Beta distribution
Beta prior distribution – $P(\theta)$

\[ P(\theta) = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T) \]

- **Likelihood function:** $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H} (1-\theta)^{\alpha_T}$
- **Posterior:** $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta)$
Posterior distribution

- Prior: $\textit{Beta}(\beta_H, \beta_T)$
- Data: $\alpha_H$ heads and $\alpha_T$ tails
- Posterior distribution:

$$P(\theta \mid \mathcal{D}) \sim \textit{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$
Using Bayesian posterior

- **Posterior distribution:**
  \[ P(\theta \mid \mathcal{D}) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

- **Bayesian inference:**
  - Estimate mean
    \[ E[\theta] = \int_0^1 \theta P(\theta \mid \mathcal{D}) d\theta \]
  - Estimate arbitrary function \( f \)
    \[ E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta \]
  - For arbitrary \( f \) integral is often hard to compute
MAP: Maximum a posteriori approximation

\[ P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

\[ E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta \]

- As more data is observed, Beta is more certain

- MAP: use most likely parameter:

\[ \hat{\theta} = \arg \max_\theta P(\theta \mid \mathcal{D}) \quad E[f(\theta)] \approx f(\hat{\theta}) \]
MAP for Beta distribution

\[ P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H + \alpha_H - 1} (1 - \theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

- MAP: use most likely parameter:

\[ \hat{\theta} = \arg \max_\theta P(\theta \mid \mathcal{D}) = \]
MAP for Beta distribution

\[
P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H + \alpha_H - 1}(1 - \theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)
\]

- MAP: use most likely parameter:

\[
\hat{\theta} = \arg \max_{\theta} P(\theta \mid \mathcal{D}) = \frac{\beta_H + \alpha_H - 1}{\beta_H + \beta_T + \alpha_H + \alpha_T - 2}
\]

- Beta prior equivalent to extra coin flips
- As \( N \to 1 \), prior is “forgotten”
- But, for small sample size, prior is important!
Bayesian vs Frequentist

- Data: $D$  
  Estimator: $\hat{\theta} = t(D)$  
  loss: $\ell(t(D), \theta)$

- Frequentists treat unknown $\theta$ as fixed and the data $D$ as random.

- Bayesian treat the data $D$ as fixed and the unknown $\theta$ as random
Recap for Bayesian learning

Bayesians are optimists:
- “If we model it correctly, we output most likely answer”
- Assumes one can accurately model:
  - Observations and link to unknown parameter $\theta$: $p(x|\theta)$
  - Distribution, structure of unknown $\theta$: $p(\theta)$

Frequentist are pessimists:
- “All models are wrong, prove to me your estimate is good”
- Makes very few assumptions, e.g. $E[X^2] < \infty$ and constructs an estimator (e.g., median of means of disjoint subsets of data)
- Must analyze each estimate