CPSC 440/540 Machine Learning - Sample Lectures

Hypothesis Testing

Outline

- Hypothesis Testing
- Neyman-Pearson
- Bayes
- MAP
- Maximum Likelihood

Outline

- **Hypothesis Testing**
- Neyman-Pearson
- Bayes
- MAP
- Maximum Likelihood

Definition

$H \in \{0, 1, \ldots, M-1\}$ represents one of M possible "states of nature". We observe a random variable Y whose distribution depends on which state of nature H is true.

The problem is to infer H from the observed signal Y . We express the relationship between H and Y by defining the likelihood function for Y for each of the M hypotheses:

$$H_0 : Y \sim p(y|H = 0)$$
$$\ldots$$
$$H_{M-1} : Y \sim p(y|H = M-1)$$

where each $p(y|H = i)$ is a density or mass function. That is, each hypothesis H_i is true if $H = i$, and for each hypothesis there is a different distribution on the signal Y .

The goal is to construct a decision rule/detector/test, i.e., a mapping $h: \mathbb{R}^N \rightarrow \{H_0, H_1, \ldots, H_{M-1}\}$ assigning an observation to a hypothesis.

Binary Hypothesis Testing

- A **null hypothesis H_0** is a statement of the status quo, one of no difference or no effect. If the null hypothesis is not rejected, no changes will be made.
- An **alternative hypothesis H_1** is one in which some difference or effect is expected.

Object detection example: A pulsed radar signal in search of a potential object.

H_0: no object -> no return pulse, only noise

H_1: object present -> return pulse + noise

We model these scenarios by univariate Gaussian likelihood functions:

$$p(y|H=0) = \mathcal{N}(0, \sigma^2)$$
$$p(y|H=0) = \mathcal{N}(a, \sigma^2)$$

Where σ^2 represents noise power and a represents amplitude of returned radar pulse.

Type I and Type II Errors

Null hypothesis is…	True	False
Rejected	**Type I Error** False positive	Correct decision True positive
Not rejected	Correct decision True negative	**Type II Error** False negative

Type I error: rejecting H_0 when H_0 is true

- α = P(type I error) is false-alarm probability, which is $P_F(\delta)=P_r(\delta(Y)=1|H=0)$, also called significance level, α-error, or **size** of the test.

Type II error: not rejecting H_0 when H_0 is false

- β = P(type II error) is missed detection probability, which is $P_M(\delta)=P_r(\delta(Y)=0|H=1)$.

Detection probability

- $P_D(\delta)=1-P_M(\delta)=P_r(\delta(Y)=1|H=1)$, also called **power** of the test.

It is necessary to balance the two types of errors.

Error

In the machine learning literature, the terms size and power are used less frequently in favor of descriptions of the precision and recall of a classifier.

- Precision

Precision measures the ratio of detections that are correct, which corresponds to (but is not equal to!) the size of the test.

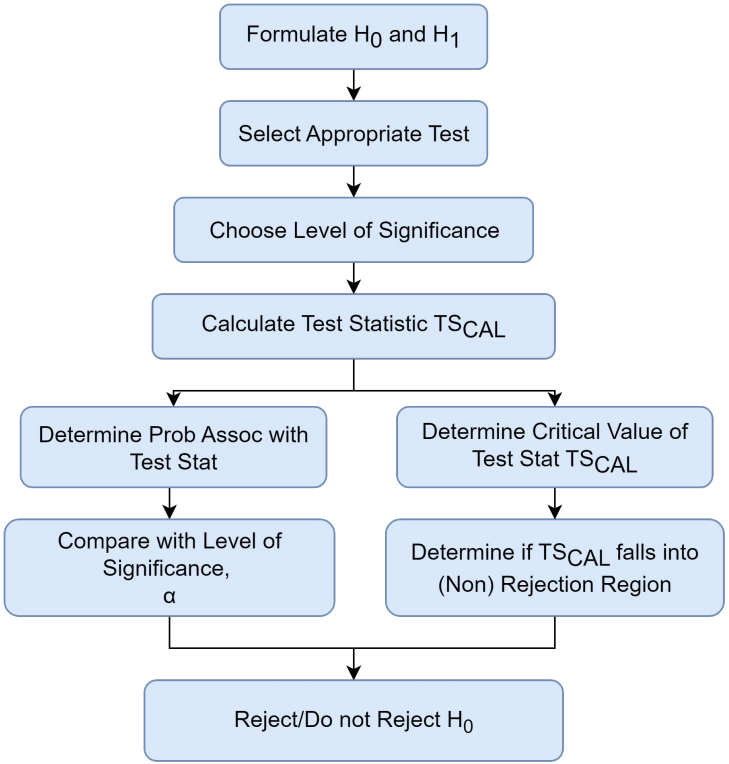
- Recall

Recall measures the fraction of H_1 events that are detected, which is essentially to equivalent to the power of the test.

- Receiver operator characteristic (ROC)

A graphical plot that shows the performance of a binary classifier system as the discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

Steps for Hypothesis Testing



Hypothesis Testing Applications

- One observation
- Multiple observations
- Spam detection
- Signal detection
- Criminal justice
- Marketing research
- Quality control
- …

Outline

- Hypothesis Testing
- **Neyman-Pearson**
- Bayes
- MAP
- Maximum Likelihood

Neyman-Pearson Criterion

For sample space Γ, let the sets Γ0 and Γ1 be a partition of Γ, we first define the decision rule δ:

$$\delta(y) = \begin{cases} 0 & if\ y \in \Gamma_0 \\ 1 & if\ y \in \Gamma_1 \end{cases}$$

A detector simply divides the set of possible values of Y (denoted as Γ) into two regions based on whether it decides that H_0 or H_1 is true. Neyman-Pearson fixes the size (denoted as α) of the test and find the most powerful detector.

$$\delta^* = \arg\max_{\delta} P_D(\delta),\ \text{subject to}\ P_F(\delta) \leq \alpha$$

We choose to tolerate a certain probability of false alarm, and we find the detector that maximizes the probability of successful detection. No detector can achieve a better trade-off than the one satisfying the Neyman-Pearson criterion.

Steps:

- Start by picking an α.
- For any α, there is an infinite number of possible decision rules (infinite number of critical regions).
- Each critical region has a power.
- Neyman Pearson Lemma tells us how to find the critical region (i.e. test) that has the highest power.

Likelihood Ratio Test

Neyman-Pearson criterion is a constrained optimization problem. The Likelihood Ratio Test (LRT) is a popular method used within the Neyman-Pearson framework to determine the optimum decision criterion.

The LRT involves computing the ratio of the likelihood of the observed data under the alternative hypothesis to the likelihood of the observed data under the null hypothesis.

Likelihood ratio: $$l(y) := \frac{p(y|H=1)}{p(y|H=0)}$$

- We then compare this ratio to a threshold value (known as the threshold likelihood ratio or decision threshold) to make a decision.
- If the ratio is greater than the threshold, we reject the null hypothesis in favor of the alternative hypothesis. Otherwise, we fail to reject the null hypothesis.

Log-likelihood Ratio Test

The Neyman-Pearson lemma shows that the likelihood ratio test (LRT) is the most powerful test of H_0 against H_1:

$$\delta(y) = \begin{cases} 0 & if\, l(y) < \eta \\ \gamma(y) & if\, l(y) = \eta \\ 1 & if\, l(y) > \eta \end{cases}$$

- The threshold of the LRT is determined by the size α of the test. The larger the size α, the more tolerant we are to Type I errors, and the lower the threshold on the likelihood ratio.
- If l(y) is larger than this threshold, we reject the H_0. Otherwise, we fail to reject H_0.
- **The log-likelihood ratio (LLR)** is frequently used to express the Likelihood Ratio Test (LRT) since many distributions involve exponential functions. Hence, we often convert the likelihood ratio to its logarithmic form. The LLR is defined as:

$$L(y) := log(l(y)) = log(p(y|H = 1)) - log(p(y|H = 0))$$

LRT $l(y) \lessgtr \eta$ can be expressed as a test on the LLR, or $L(y) \lessgtr v$, where $v := \log \eta$ because logarithm is monotonically increasing.

Neyman Pearson Example

Object Detection example continued:

$$p(y|H=0) = \mathcal{N}(0, \sigma^2)$$
$$p(y|H=0) = \mathcal{N}(a, \sigma^2)$$

The log likelihood ratio is:

$$L(y) = \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(y-a)^2}{2\sigma^2} - \log \frac{1}{\sqrt{2\pi\sigma^2}} + \frac{y^2}{2\sigma^2}$$
$$= \frac{a^2 - 2ay}{2\sigma^2}.$$

The Neyman-Pearson detector is therefore:

$$\delta(y) = \begin{cases} 0 & \text{if } y \leq \frac{\sigma^2}{a}\nu + \frac{a}{2} \\ 1 & \text{if } y > \frac{\sigma^2}{a}\nu + \frac{a}{2} \end{cases}.$$

For convenience, define: $y^* := \frac{\sigma^2}{a}\nu + \frac{a}{2}$

The choice of **v** (or η or y*) is dependent on each test. For this example, we compute the size of the test as a function of the threshold value y*:

$$P_F(\delta) = \int_{y^*}^{\infty} p(y|H=0)dy = 1 - \Phi\left(\frac{y^*}{\sigma}\right)$$

Therefore, we choose y* based on a test of size α as:

$$y^* = \sigma\Phi^{-1}(1-\alpha)$$

where Φ^{-1} is the inverse CDF of the normal distribution.

Neyman-Pearson ROC

We can use the ROC curve to identify the decision threshold that achieves the desired balance between P_D and P_F. We usually plot α on the x-axis and $P_D(\delta)$ of the resulting detector on the y-axis.

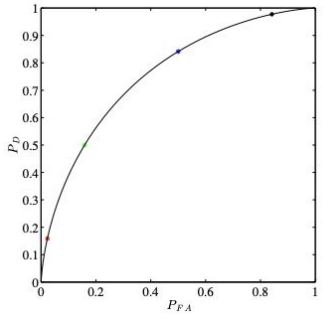


Figure 1. A typical ROC Curve

It shows the tradeoff between the false alarm probability $P_F(\delta)$ and the detection probability $P_D(\delta)$ of the Neyman-Pearson detector δ associated with every $P_F(\delta) = \alpha$.

- Starts from (0, 0) and ends at (1, 1) in most cases unless $p_i(\pm\infty) > 0$. Achieving zero false alarms would mean that we have to accept zero correct detections, while allowing all false alarms means we will never miss a detection.
- The diagonal line from (0, 0) to (1, 1) corresponds to random guesses.
- Depends on signal strength, noise strength, noise type, etc.

Neyman-Pearson ROC

Object Detection example continued: $p(y|H=0) = \mathcal{N}(0, \sigma^2)$

$p(y|H=0) = \mathcal{N}(a, \sigma^2)$

We previously computed the size of the test as: $P_F(\delta) = \int_{y^*}^{\infty} p(y|H=0)dy = 1 - \Phi\left(\frac{y^*}{\sigma}\right)$

And chose y* as: $y^* = \sigma\Phi^{-1}(1-\alpha)$

Now compute the detection probability: $P_D(\delta) = \int_{y^*}^{\infty} p(y|H=1)dy = 1 - \Phi\left(\frac{y^*-a}{\sigma}\right)$

Substitute y* and get the trade-off: $P_D(\alpha) = 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{a}{\sigma}\right)$

Plot the ROC curve for chosen values of a and σ^2:

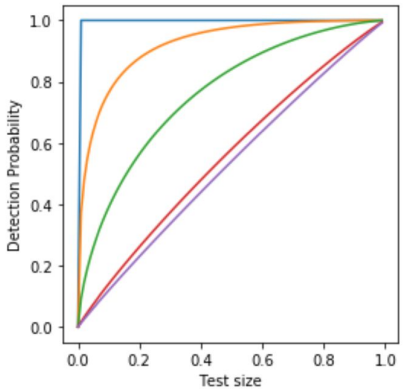


Figure 2. ROC with a = 1 and $\sigma^2 \in \{0.1, 0.5, 1, 5, 10\}$ *(Nokleby, 2018)*

Outline

- Hypothesis Testing
- Neyman-Pearson
- **Bayes**
- MAP
- Maximum Likelihood

Prior Distribution

- Different from NP Hypothesis Testing, Bayes Hypothesis Testing we also assume we have **some knowledge** on the distribution of each hypothesis p(H).
- It is called prior distribution, binary cases:

$$\pi_0 = p(H = 0) = p_0$$
$$\pi_1 = p(H = 1) = p_1$$

- Sum to 1 property (discrete): $\sum_{h=1}^{H} \pi_h = 1$
- Binary case (know one = know all): $\pi_0 = 1 - \pi_1$

Bayesian Detector (Binary Case) -- Error Rate

- Prior: two cases
- Detection error (E): need to consider two cases

$$E = \begin{cases} 0 & \text{if } \delta(Y) = H \\ 1 & \text{if } \delta(Y) \neq H \end{cases}$$

- The probability of an error happens (conditional probability):

$$p(E = 1|H = 1) = \int_{\Gamma_0} p(y|H = 1)dy$$
$$p(E = 1|H = 0) = \int_{\Gamma_1} p(y|H = 0)dy$$

- Overall error rate (marginal rule):

$$\pi_0 \int_{\Gamma_1} p(y|H = 0)dy + \pi_1 \int_{\Gamma_0} p(y|H = 1)dy$$

Guess 0, but wrong

Guess 1, but wrong

Bayesian Detector (Binary Case) -- Detector Formation

- Tool: Likelihood Ratio Test
- Comparison: which H more likely goes wrong?

$$\pi_0 \int_{\Gamma_1} p(y|H=0)dy + \pi_1 \int_{\Gamma_0} p(y|H=1)dy$$

Guess 0, but wrong | Guess 1, but wrong

- red > blue: less likely go wrong if pick H=1
- blue > red: less likely go wrong if pick H=0
- Detector Form:

$$\delta(y) = \begin{cases} 0, & \text{if } \frac{p(y|H=1)}{p(y|H=0)} < \frac{\pi_0}{\pi_1} \\ 1, & \text{if } \frac{p(y|H=1)}{p(y|H=0} \geq \frac{\pi_0}{\pi_1} \end{cases}$$

What if the priors are equal?

- likelihood ratio test

What if we replace prior ratio with a constant η?

- Go back to Neyman-Pearson
  - $\pi_o \to 0$, $\eta \to 0$, $\delta(y) = 1$
  - $\pi_o \to 1$, $\eta \to \infty$, $\delta(y) = 0$

Risk

- Recall from CPSC 340 -- cost in bayes spam detector [https://www.cs.ubc.ca/~schmidtm/Courses/340-F22/L6.pdf]

  We can give a cost to each scenario, such as:

Predict / True	True 'spam'	True 'not spam'
Predict 'spam'	0	100
Predict 'not spam'	10	0

  - Overall Cost: Weighted sum of conditional probabilities
- Bayes hypothesis testing has similar but more general, named as **risk**
- Binary cases

Cost function	H_0	H_1
H_0	C_{00}	C_{01}
H_1	C_{10}	C_{11}

If C_{10} is substantially larger than C_{01}, we take care to declare H_1 as the hypothesis only when we are very sure that we are correct!

Risk

- Risk for the Bayes Detector: $r(\delta)$

$$r(\delta) = C_{00}P(H_0, \delta(Y) = 0) + C_{10}P(H_0, \delta(Y) = 1) + C_{01}P(H_1, \delta(Y) = 0) + C_{11}P(H_1, \delta(Y) = 1)$$

P(A,B) = P(A|B)P(B)

$$= \int_{\Gamma_0} (C_{00}\pi_0 p(y|H_0) + C_{01}\pi_1 p(y|H_1))dy + \int_{\Gamma_1} (C_{10}\pi_0 p(y|H_0)dy + C_{11}\pi_1 p(y|H_1))dy.$$

- Γ_0: Y = 0 is true; Γ_1: Y = 1 is true;
- Green: H is correct; red: H is incorrect
- Choose either hypothesis can be right or wrong
- Typically, C_{00} and C_{11} are low (guess correct)
- If C10 is substantially larger than C01, we take care to declare H1 as the hypothesis only when we are very sure that we are correct!

Minimize Bayes Risk

- Start from a simple case:
  - $C_{00} = C_{11} = 0$, $C_{01} = C_{10} = 1$
  - Minimizing risk = minimizing the error probability
    - Using LRT
- General case:
  - $C_{00} \neq C_{11} \neq C_{01} \neq C_{10}$
  - For each y, the risk can be expressed as:
  - $r(\delta(y)) = C_{00}\pi_0 P(y|H=0) + C_{01}\pi_1 P(y|H=1) + C_{10}\pi_0 P(y|H=0) + C_{11}\pi_1 P(y|H=1)$
  - Red: risk of detect as H=0; Blue: risk of detect as H=1
  - Minimizing the risk = minimizing the risk of individual y = pick the box with the lowest value of risk
  - Example: when $\delta(y) = 1 \rightarrow$ lower risk

$$C_{00}\pi_0 P(y|H=0) + C_{01}\pi_1 P(y|H=1) > C_{10}\pi_0 P(y|H=0) + C_{11}\pi_1 P(y|H=1)$$
$$C_{00}\pi_0 P(y|H=0) - C_{10}\pi_0 P(y|H=0) > C_{11}\pi_1 P(y|H=1) - C_{01}\pi_1 P(y|H=1)$$
$$\pi_0 P(y|H=0)(C_{00} - C_{10}) > \pi_1 P(y|H=1)(C_{11} - C_{01})$$
$$\frac{P(y|H=0)}{P(y|H=1)} > \frac{\pi_1(C_{11} - C_{01})}{\pi_0(C_{00} - C_{10})}$$

Practice: build the full detector

Bayesian Multiple Hypothesis Testing

- Let's go more general
- Definition:
  - M hypothesis {0, 1, 2, …, M-1}
    - $H_i = P(y|H = i)$
  - M priors $\{\pi_0, \pi_1, \ldots, \pi_{M-1}\}$ (Marginal: all priors sum to 1)
  - $\delta(y)$ can be M outcomes
  - Cost: $C_{ij} \rightarrow \delta(y) = i$ and $H = j$
- Goal: build a detector with the **minimum** risk

$$
\begin{aligned}
r(\delta) &= \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} C_{ij} \Pr(\delta(Y) = y, H = j) \\
&= \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} \pi_j \int_{\Gamma_i} p(y|H = j) dy, \\
&= \sum_{i=0}^{M-1} \int_{\Gamma_i} \sum_{j=1}^{M-1} C_{ij} \pi_j p(y|H = j) dy
\end{aligned}
$$

Overall risk: iterate over label and hypothesis set, and sum

Bayesian Multiple Hypothesis Testing

- Best detector: the detected hypothesis H has the **lowest** risk compared with the remaining hypothesis set
  - risk for single case i:

$$r(\delta(y) = i) = \sum_{j=1}^{M-1} C_{ij}\pi_j p(y \mid H = j)$$

  - use argmin to locate the detection result with given y

$$\delta(y) = \arg\min_i \sum_{j=1}^{M-1} C_{ij}\pi_j p(y|H = j)$$

Outline

- Hypothesis Testing
- Neyman-Pearson
- Bayes
- **MAP**
- Maximum Likelihood

MAP - Special Case with Risk

- 0 cost if correct, cost = 1 if incorrect

$$C_{ij} = \begin{cases} 1, & if\ i \neq j \\ 0, & if\ i = j \end{cases}$$

- remove the case that prediction is correct
- New risk function:

$$r(\delta(y) = i) = \sum_{j=1, i \neq j}^{M-1} \pi_j p(y \mid H = j)$$

- New detector: argmin(new risk function)
  - Can we come up with a more meaningful expression?

MAP - Special Case with Risk

- Can we come up with a more meaningful expression?

$$\delta(y) = \arg\min_i \left(\sum_{j=0, j\neq i}^{M-1} \pi_j p(y \mid H = j) \right)$$

$$= \arg\min_i \left(\sum_{j=0}^{M-1} \pi_j p(y \mid H = j) - \pi_i p(y \mid H = i) \right)$$

independent with argmin function

$$= \arg\max_i \left(\pi_i p(y \mid H = i) \right)$$

$$\propto \arg\max_i \left(p(H = i|y) \right)$$

P(B)P(A|B) = P(B|A)/P(A)

- 0 cost if correct, cost = 1 if incorrect → MAP

Outline

- Hypothesis Testing
- Neyman-Pearson
- Bayes
- MAP
- **Maximum Likelihood**

Maximum Likelihood Multiple Hypothesis Testing

- Another method to select one hypothesis from multiple candidates.

$$\delta(y) = \arg\max_i p(y|H = i)$$

- Computing the likelihood of observed data
- In Bayes Hypothesis Testing
  - prior is uniform: special case for MAP
  - cost: same as MAP
- Used for parameter estimation (Covered in CPSC 440)

Summary

- Hypothesis testing can be used in signal detection, spam detection, quality control…
- Neyman-Pearson method is binary hypothesis testing method that utilizes likelihood ratio test to maximize the detection probability while restraining the false alarm rate within p-value.
- Bayesian method can be used in binary or multiple hypothesis testing, it builds the detector based on minimizing the risk.
- MAP and MLE are special cases of Bayesian methods.

Summary/ Remarks

- Detection, such as Neyman-Pearson, is based on the distribution generated from signals to features, and selects a solution from a discrete hypothesis space; classification, such as linear discriminant analysis, is fitting the model with data without knowing the distribution and makes discrete predictions.
- The slides follow the conventions in detection to indicate **observations** as **Y**. To make a fair comparison with **features** in machine learning methods, the assignments use **X** to indicate the **observations**.

References

Nokleby, M. (2018). *Detection, Estimation, and Learning*. detection-estimation-learning. Retrieved April 21, 2023, from https://github.com/docnok/detection-estimation-learning/blob/master/course-notes.pdf

Maximum likelihood - psychology and Neuroscience. (n.d.). Retrieved April 27, 2023, from http://psych.colorado.edu/~carey/qmin/qminChapters/QMINA2-Maximum_Likelihood.pdf

CPSC 340: Machine Learning and Data Mining Non-Parametric Models Fall 2022. (n.d.). Retrieved April 27, 2023, from https://www.cs.ubc.ca/~schmidtm/Courses/340-F22/L6.pdf