

Overview

- Last time:
 - decision trees
 - started maximum likelihood (ML) hypothesis testing
- This time:
 - continue ML hypothesis testing
 - limitations of ML hypothesis testing

Hypothesis Testing

- Let us now consider cases when we cannot always make decisions with certainty
- We want to test a hypothesis:
 - whether or not my radar detected a target
 - whether or not a chip is good
 - whether or not a new treatment is effective.

Maximum Likelihood

- Pick the one that maximizes the probability of what we observed
- Hypotheses $H_1, H_2, H_3, \dots$
- We observe A_j
- We select the H_i for which $P(A_j)$ is highest
- We have different probability measures to use for each H_i

Communications example

- Two users, **User 1** and **User 2**, communicate, each differentiated by a code
- **User 1: 1000**
- **User 2: 0001**
- Consider that **User 2** sent but that 2 bits get flipped: the first and the last bit, i.e. Output is **1000**
- ML true hypothesis is **User 1**

In terms of Conditional Probabilities

- We look at the H_i maximizing $P(A_j | H_i)$
- We call H_i the true hypothesis
- How would we calculate our probability of error?

$$\sum_{\text{all } H_m} \frac{P(H_m)}{P(A_j | H_m)} \sum_{\text{all } A_j \text{ such that } P(A_j | H_k) > P(A_j | H_m)} P(A_j | H_m)$$

Conditional Probabilities

	A_1	A_2								
User 1	1	0	1	0	1	1	1	1	0	0
	0	0	1	0	1	0	0	0	1	1
	0	0	1	1	0	1	0	1	0	1
	0	1	0	1	1	1	1	0	0	0

$$\sum_{\text{all } H_m} P(H_m) \sum_{\text{all } A_j \text{ such that } P(A_j | H_k) > P(A_j | H_m) \text{ for some } k} P(A_j | H_m)$$

$$= P(\text{User 1}) \times [P(A_2 | \text{User 1}) + P(A_4 | \text{User 1}) \dots]$$

$$+ P(\text{User 2}) \times [P(A_1 | \text{User 2}) + P(A_3 | \text{User 2}) + \dots]$$

User 2	1	0	1	0	1	1	1	1	0	0
	0	0	1	0	1	0	0	0	1	1
	0	0	1	1	0	1	0	1	0	1
	0	1	0	1	1	1	1	0	0	0

Think - Pair Share

- Suppose that the probability of any particular bit being flipped is $1/8$
- What is the ML true hypothesis for 1100?
- What is the ML true hypothesis for 0101?
- What is the ML true hypothesis for 1001?

Think - Pair Share

- Suppose that the probability of any particular bit being flipped is $1/8$
- What is the ML true hypothesis for 1100?
- **User 1**
- What is the ML true hypothesis for 0101?
- **User 2**
- What is the ML true hypothesis for 1001?
- **Cannot discriminate!**

Example

- A target is either present or absent for a radar measurement
- The reading is HI when the target is present 95 % of the time
- The reading is HI when the target is not present 20 % of the time.

Think - Pair Share

- Do we know the probability that the reading is LO when the target is absent?
- Do we know the probability that the reading is LO when the target is present?

Think - Pair Share

- Do we know the probability that the reading is LO when the target is absent?
- **80 %**
- Do we know the probability that the reading is LO when the target is present?
- **5 %.**

Self-study: figure out probability of error is $P(\text{target}) = 0.1$, $P(\text{no target}) = 0.9$
Read class notes pp. 25-30.

What are the weaknesses of ML?

- Consider our communications example
- Suppose that **User 1** is a priori ninety times more likely to be present than **User 2**
- $P(\text{User 1}) = 0.99$, $P(\text{User 2}) = 0.01$
- What is the ML true hypothesis for 1100?
- What is the ML true hypothesis for 1001?

What are the weaknesses of ML?

- Consider our communications example
- Suppose that **User 1** is a priori 99 times more likely to be present than **User 2**
- $P(\text{User 1}) = 0.99$, $P(\text{User 2}) = 0.01$
- What is the ML true hypothesis for 1100?
- What is the ML true hypothesis for 1001?
- **We cannot distinguish** - yet something tells us that in some sense **User 1** is more likely!

Let us look at the Probability of Error

- What are the terms associated with 1001?

Let us look at the Probability of Error

- What are the terms associated with 1001?
- If we decide that inability to distinguish is an error:
- $P(\text{User 1}) \times P(1001 | \text{User 1}) + P(\text{User 2}) \times P(1001 | \text{User 2}) = 1/8 \times (7/8)^3$
- It does not depend on $P(\text{User 1})$, $P(\text{User 2})$

Let us look at the Probability of Error

- What are the terms associated with 1001?
- If we decide to pick **User 1**
- $P(\text{User 2}) \times P(1001 | \text{User 2}) = 0.01 \times 1/8 \times (7/8)^3$
- If we decide to pick **User 2**
- $P(\text{User 1}) \times P(1001 | \text{User 1}) = 0.99 \times 1/8 \times (7/8)^3$
- It **does** depend on $P(\text{User 1})$, $P(\text{User 2})$

Let us look at the Probability of Error

- How about 0011?
- What is the ML true hypothesis?

Let us look at the Probability of Error

- How about 0011?
- What is the ML true hypothesis? **User 2!**

Let us look at the Probability of Error

- How about 0001?
- What is the ML true hypothesis? **User 2!**
- What is the term in the probability of error which is associated with 0001?
- We decide to pick **User 2**
- $P(\text{User 1}) \times P(0001 | \text{User 1}) = 0.99 \times 1/64 \times (7/8)^2 = 0.011843$

Let us look at the Probability of Error

- For 0001: ML true hypothesis **User 2**
- **WHAT IF WE DECIDE TO PICK USER 1 INSTEAD?**
- What is the term in the probability of error which is associated with 0001?
- We decide to pick **User 1**
- $P(\text{User 2}) \times P(0001 | \text{User 2}) = 0.01 \times (7/8)^4 = 0.0059$

What's going on?

- Our probability of error is lower by making a decision which is not the ML true hypothesis
- Can we capture this in some different way?
- Let us try looking at $P(\text{User } i | 0001)$
- **MAXIMUM A POSTERIORI**
- Use Bayes
- $P(\text{User } i \text{ and } 0001) / (P(\text{User 1}) \times P(0001 | \text{User 1}) + P(\text{User 2}) \times P(0001 | \text{User 2}))$

Think - Pair Share

- WITHOUT USING a calculator, say which decision you would make from observing 0001

Think - Pair Share

- WITHOUT USING a calculator, say which decision you would make from observing 0001
- **User 1**

Reality Check

- Does this make sense?
- Are we always going to pick **User 1** anyway, since even **User 2's** code maps onto **User 1** with this method of making decisions?

Reality Check

- Does this make sense?
- Are we always going to pick **User 1** anyway, since even **User 2's** code maps onto **User 1** with this method of making decisions?
- Picking **User 1** just makes more sense for 0001. What about 0011?
- How could we remedy this?

Why would we ever use ML?

- We don't always have the a priori probabilities
- Example: **parameter estimation**.
- Remember: random variable is a real associated with the elements of some Ω
- Other way of looking at it: r.v. is a function, or mapping that maps Ω into real numbers

How do we determine the probability of a random variable?

- Example: tossing a coin with probability p of Heads
- This is a Bernoulli random variable:
 - $P(\text{Heads}) = P(X=1) = p$
 - $P(\text{Tails}) = P(X=0) = 1-p$
- What is the probability that we get exactly i Heads when we toss the coin n times? Let Z be the r.v. which takes the value i if we got i Heads when we tossed the coin n times.

Binomial Random Variable

$$P(Z=i) = \binom{n}{i} p^i (1-p)^{n-i}$$

Z is a binomial random variable with parameters (n,p)

Verify: take $n=5$, $i=2$, the probability of HHTTT is $p^2 (1-p)^3$ same for any combination with 2 H and 3 T and the number of such combinations is:

$$\binom{5}{2}$$

Check

- Do the probabilities add to 1?
- Use the Binomial Theorem!

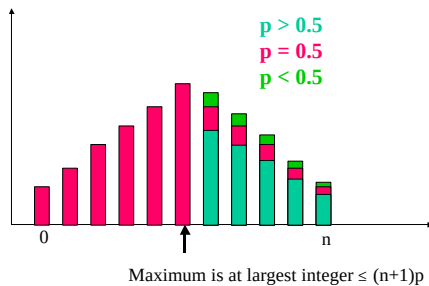
$$\sum_{i=0}^n P(Z=i) = \sum_{i=0}^n \binom{n}{i} p^i (1-p)^{n-i}$$

Self-study: Read Ross 4.7

What if I am trying to determine parameter p ?

- I get to toss a coin several times to try to determine its p
- We observe the incidence of H and T
- ML estimation: do not need a probability for the different values of p
- We observe k Heads in n tosses - pick the p which maximizes $P(Z=k)$ where Z is a binomial (n, p) r.v.

Binomial Probability Mass Function (p.m.f)



We observe it - but can we show it?

$$\frac{P(Z = k)}{P(Z = k - 1)} = \frac{\frac{n!}{(n-k)!k!} p^k (1-p)^{n-k}}{\frac{n!}{(n-k+1)!(k-1)!} p^{k-1} (1-p)^{n-k+1}}$$

$$= \frac{p(n-k+1)}{k(1-p)}$$

Prop. 7.1 on pg. 150 in Ross.

So how would we do ML estimation of p?

- I toss a coin 100 times
- I get exactly 20 H
- What is our intuitive guess?

So how would we do ML estimation of p?

- I toss a coin 100 times
- I get exactly 20 H
- What is our intuitive guess?
- **P = 1/5!**
- Let's check it mathematically.

$$P(Z = k) = \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k}$$

$$\frac{\partial P(Z = k)}{\partial p} = \frac{n!k}{(n-k)!k!} p^{k-1} (1-p)^{n-k} - \frac{n!(n-k)}{(n-k)!k!} p^k (1-p)^{n-k-1}$$

$$\frac{\partial P(Z = k)}{\partial p} = 0 \Leftrightarrow \frac{n!k}{(n-k)!k!} p^{k-1} (1-p)^{n-k} - \frac{n!(n-k)}{(n-k)!k!} p^k (1-p)^{n-k-1} = 0 \Leftrightarrow p = k$$

MAXIMIZE the probability of what we observed

Let us look at another type of r.v.

- The Poisson r.v., let us call it X
- we use it to model arrivals to queues, births in a certain population size in an interval of time t
- more of it anon when we look at independence

$$P(X = i) = e^{-t} \frac{(t)^i}{i!}$$

Taylor series expansion

$$\sum_{i=0}^{\infty} P(X = i) = \sum_{i=0}^{\infty} e^{-t} \frac{(t)^i}{i!} = e^{-t} \sum_{i=0}^{\infty} \frac{(t)^i}{i!} = 1$$

Note : if no t is given, then it is understood that t=1 and λ is the only parameter

Read 4.8 in Ross.

ML estimation of λ

- If we observe that $X=i$, then we maximize that probability with respect to λ

$$\frac{\partial P(X=i)}{\partial l} = -e^{-l} \frac{l^i}{i!} + e^{-l} i \frac{l^{i-1}}{i!} = 0$$
$$\Leftrightarrow l = i$$