香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

Introduction to Data Science

Lecture 10 Statistics

Maximum Likelihood Estimate

Zicheng Wang

Review

- **Elementary Probability Theory**
  - Sample space, independence, mean/variance, PDF, CDF, etc.
    - Most important
  - Common probability mass/density functions.
    - No need to know other distributions in our course.

Bonus Question

- Two Blind Red Envelops
- The money inside is independently drawn from a uniform $[0,100]$ distribution
- You select envelope one, but are allowed to switch to envelope two upon seeing the realization of the first envelop
- What's best strategy to pick the one with the largest reward?

Bonus Question

Independent and identically distributed

Two envelopes have i.i.d rewards R_1, R_2 that are drawn from a uniform $[0, 1]$ distribution. Because the rewards are i.i.d, selecting any of the envelopes gives you a 50% chance of selecting the one with the largest reward. Suppose that you select envelope one, but are allowed to switch to envelope two upon seeing the realization of the first envelope. A friend of yours proposes the following switching policy: pick a number $t \in (0, 1)$. If $R_1 > t$, then keep R_1 and otherwise switch to R_2.

(a) (4 points) What is the probability of selecting the envelope with higher reward under this strategy for arbitrary t?

(b) (3 points) What would be the optimal choice of t?

(c) (3 points) What would be the expected reward under the optimal t?

Solution:

(a)

Let $\tau(t)$ be the probability of selecting the envelope with higher reward with this strategy. Then

$$\begin{aligned}
\tau(t) &= \mathbb{P}(R_2 > R_1, R_1 < t) + \mathbb{P}(R_1 > R_2, R_1 > t) \\
&= \int_0^t \int_u^1 dvdu + \int_t^1 \int_0^u dvdu \\
&= \int_0^t (1-u)du + \int_t^1 udu \\
&= t - 0.5t^2 + 0.5(1-t^2) \\
&= 0.5 + t - t^2.
\end{aligned}$$

(b)

$\tau(t)$ is a concave function that is maximized at $t^* = 0.5$, resulting in $\tau(0.5) = 3/4$.

(c)

The expected reward of this policy is

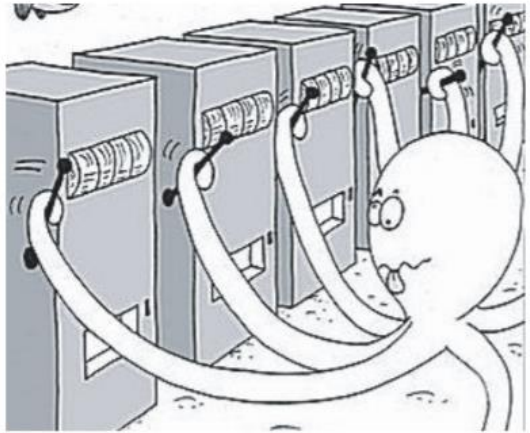
$$\begin{aligned}
&\mathbb{E}[R_1|R_1 > t]\mathbb{P}(R_1 > t) + \mathbb{E}[R_2|R_1 \leq t]\mathbb{P}(R_1 \leq t) \\
&= \int_t^1 udu + E[R_2]t \\
&= 0.5(1-t^2) + 0.5t \\
&= 0.5(1 + t - t^2).
\end{aligned}$$

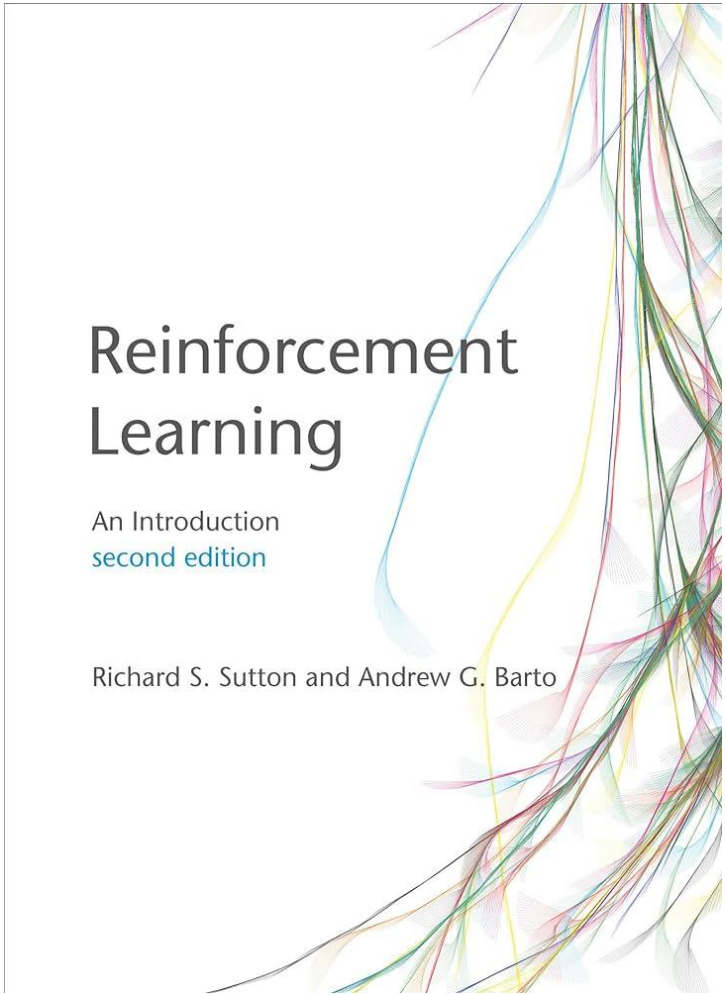
So at $t^* = 1/2$ the expected reward is $5/8$.

Bonus Question

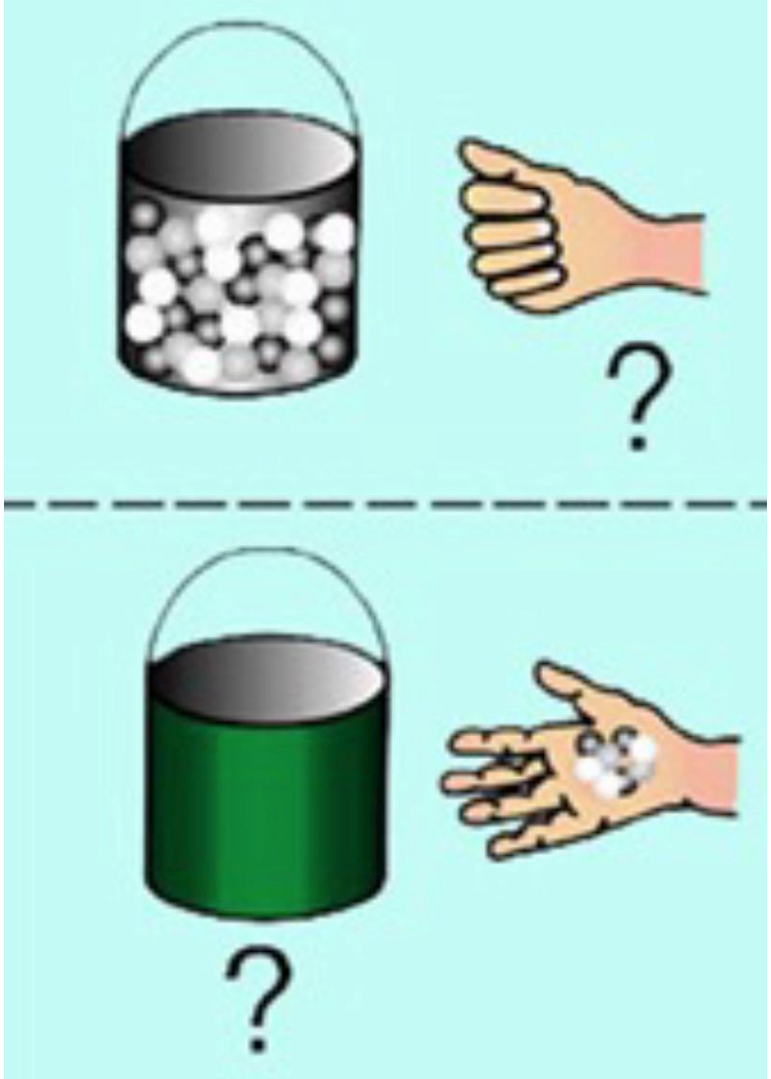
- Two Blind Red Envelops
- Suppose the reward in one envelop is drawn from a uniform $[0,100]$ distribution, the other has a uniform $[0,50]$ distribution
- You can pick one envelop every day of this year
- What should you do?

- It is a very hard problem
- Multi-armed bandit problem





- Let’s consider an easier problem
- One Blind Red Envelop
- Suppose the reward inside is either drawn from a uniform $[0,100]$ distribution, or a uniform $[0,50]$ distribution
- How can you decide which distribution it is?



Probability

- **Knowing** the true **model**, you can **predict** the **samples** outcome.

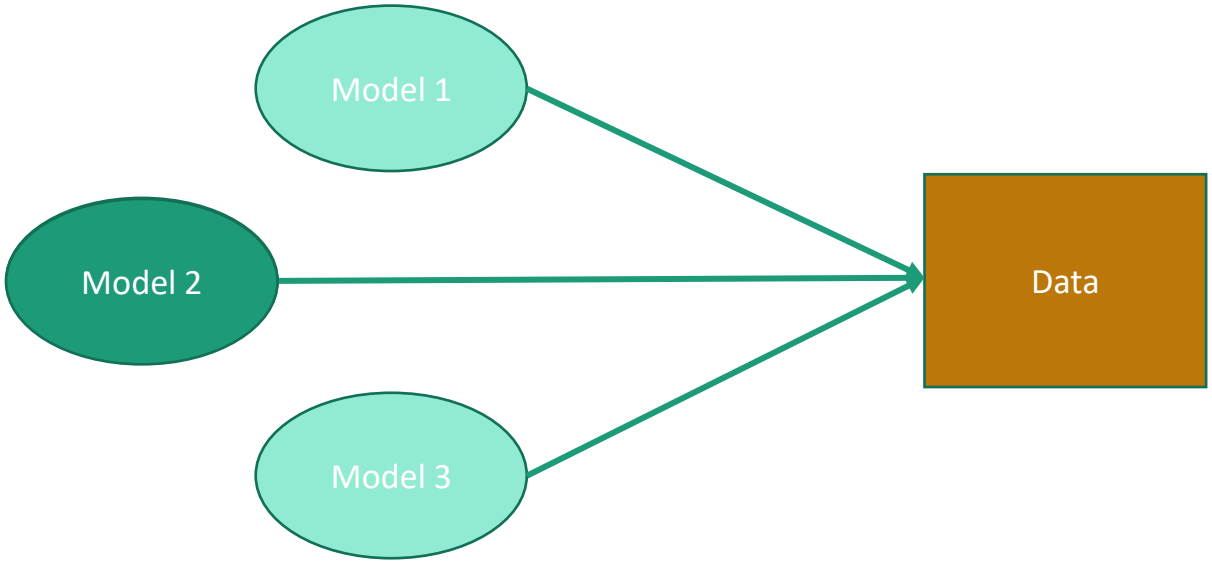
Statistics

- **Knowing** the generated **samples**, you can **estimate** the true **model**.

The general setup

Example: We have a data set, and we want to **infer** the properties of the underlying distribution from this data set.

- A variety of different probability models must at least be explored
- To see which model best “fits” the data.



In other words, you need

- Samples
- A set of possible models
- Criterion for quantifying model performance

Reading Materials

- Applied Statistics and Probability for Engineers, Third Edition, Douglas C. Montgomery and George C. Runger.

- Chap 7-3.2: Method of Maximum Likelihood

Example 1

- Whether a drug can cure a disease is a Bernoulli RV
- Yes (X=1) w.p. p; no (X = 0) w.p. 1 − p.

- p is unknown.

- 100 Samples: 81 of them were clinically successful.

- What's your estimation of p? 0.81

Example 1

- Whether a drug can cure a disease is a Bernoulli RV
- Yes (X=1) w.p. p; no (X = 0) w.p. 1 − p.

- p is unknown.

- N Samples: M of them were clinically successful.

- What's your estimation of p? M/N

Definition

- Given samples: $X_1, X_2 \ldots, X_n$
- Goal: estimate an unknown parameter θ of the true model, (we use samples).

The hat notation

The point estimator used to estimate a parameter θ is usually denoted as $\widehat{\theta}$.

Definition

- Given samples: $X_1, X_2 \ldots, X_n$
- Goal: estimate an unknown parameter θ of the true model, (we use samples).
- The estimate $\hat{\theta}$ clearly depends on $(X_1, X_2 \ldots, X_n)$, why?
  - $\hat{\theta}$ is a function of samples $(X_1, X_2 \ldots, X_n)$
  - $\hat{\theta}$ extracts some (useful) information from the samples.
- For example, you can use $\overline{\boldsymbol{X}} \coloneqq \frac{\sum_i \boldsymbol{X_i}}{\boldsymbol{N}}$ **(sample mean)** as $\hat{\theta}$

Definition

- Given samples: $X_1, X_2 \ldots, X_n$
- Goal: estimate an unknown parameter θ of the true model, (we use samples).
- The estimate $\hat{\theta}$ clearly depends on $(X_1, X_2 \ldots, X_n)$, why?
  - $\hat{\theta}$ is a function of samples $(X_1, X_2 \ldots, X_n)$
  - $\hat{\theta}$ extracts some (useful) information from the samples.
- And we call $\hat{\theta}$ the **statistic.**

How to choose statistic?

- Whether a drug can cure a disease is a Bernoulli RV
- Yes (X=1) w.p. p; no (X = 0) w.p. 1 − p.

- N Samples: M of them were clinically successful.
- What's your estimation of p? M/N

Why you use $\frac{M}{N}$ to estimate p?

Example 2

- Drug 1: p = 0.8
- Drug 2: p = 0.9
- Samples: $\mathrm{M} = 81$ successes from N=100 experiments.
- Which drug do you think is experimented?

How to determine the best parameter?

- Among **those possible choices**, choose the one that is **most likely** to generate the samples we have.

Example 1

- For drug experiments,
  - **N** experiments
  - **M** successes

- Possible models: Bernoulli distribution with probability p.

- Given one model, the probability of generating such samples:

$$L(p) = p^{M}(1-p)^{N-M}$$

- Most likely: the model **maximizes** the probability L(p)

Example 1

- Given one model, the probability of generating such samples:

$$L(p) = p^{M}(1-p)^{N-M}$$

- Maximize the probability

We often maximize the **logarithm of** the probability (usually easier to maximize) of generating such samples

$$l(p) = \log L(p) = M * \log p + (N-M) * \log (1-p)$$

Example 1

- To maximize a continuous function l(p)
  - Find the point $l'(p) = 0$ and $l''(p)<0$
  - Before we discuss optimization, all exercises with continuous cases just require you to show the point $l'(p) = 0$.

$$l(p)=\log L(p) = M * \log p + (N-M) * \log (1-p)$$

- $l'(p)= \frac{M}{p} + \frac{N-M}{1-p} = 0$  p = M/N (sample cure rate)

Formal Definition

- Given a model with an unknown parameter θ
- Given samples: $X_1,\ X_2 \ldots,\ X_n$
- The probability that the models generates the samples is called **likelihood**. $L(\theta) = P(X_1,\ X_2 \ldots,\ X_n \mid \theta)$
- To determine the best θ, we choose $\hat{\theta}$ such that $L(\theta)$ is maximized at $\theta = \hat{\theta}$. **Maximum likelihood estimate (MLE)**

Formal Definition

- Define $l(\theta) = \log L(\theta)$, which will be called log-**likelihood.**
  - In our course, log represents natural logarithm

- MLE: choose $\hat{\theta}$ such that $l(\theta)$ is maximized at $\theta = \hat{\theta}$

Likelihood function

- Given a model with an unknown parameter θ
- Given samples: $X_1,\ X_2 \ldots,\ X_n$

Continuous RV model:

- Likelihood: L(θ)= $\prod_i$ f($X_i \mid \theta$)
- Log-Likelihood: $l(\theta)$= $\sum_i \log(\text{f}(X_i \mid \theta))$

f: the model's probability density function (PDF)

Discrete RV model:

- Likelihood: L(θ)= $\prod_i$P ($X_i \mid \theta$)
- Log-Likelihood: $l(\theta)$= $\sum_i \log(\text{P}(X_i \mid \theta))$

P: the model's probability mass function (PMF)

Likelihood function

- Given a model with an unknown parameter θ
- Given samples: $X_1,\ X_2 \ldots,\ X_n$

Continuous RV model: $l(\theta) = \sum_i \log(\mathrm{f}(X_i \mid \theta))$
Discrete RV model: $l(\theta) = \sum_i \log(\mathrm{P}(X_i \mid \theta))$

Possible models: $\theta \in \Theta$ (additional information)

- Whether Θ are discrete or continuous.

Example 3: $\Theta = [0, \infty)$

- Suppose lifetime of each bulb is an exponential distribution.
  - Given parameter λ, pdf is $f(x) = \lambda\, e^{-\lambda x}$

- Samples of lifetime: $X_1, X_2 \ldots, X_n$

- L(λ) = $\lambda^n\ e^{-\lambda \Sigma_i X_i}$
- $l(\lambda)$ =Log L(λ) = n log $\lambda\ -\ \lambda\ \Sigma_i X_i$

- $l'(\lambda)\ = \frac{n}{\lambda} - \Sigma_i X_i = 0$ ⟹ $\hat{\lambda}$ = n / $\Sigma_i X_i = 1/\bar{X}$.

Example 4: $\Theta = (-\infty, \infty)$

- Suppose heights of people follow a normal distribution.
  - Given parameter μ, the model is $N(\mu, 1)$ Mean μ, variance 1
- Samples: $X_1,\ X_2 \ldots,\ X_n$
- L(μ) = $\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_i (X_i-\mu)^2}$
- $l(\mu)$ = Log L(μ) = $-\frac{1}{2}\sum_i (X_i - \mu)^2$ - $\frac{n}{2}\log 2\pi$
- $l'(\mu)$ = $\sum_i (X_i - \mu)$= 0 ⟹ $\hat{\mu} = \bar{X}$.

Example 5: discrete Θ

- You may have additional information.
  - Drug 1: p = 0.8
  - Drug 2: p = 0.9
  - Samples: M=81 successes from $N = 100$ experiments.

- Potential models: Bernoulli with $p \in \{0.8, 0.9\}$

l = Log L = M*log p+(N-M)*log (1-p)

- Drug 1: -48.6539
- Drug 2: -52.2833

Conclusion: drug 1 is more likely to be experimented.