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

Introduction to Data Science

Lecture 13 Review

Zicheng Wang

Probability

- Sample Space, Events,….
- PMF, PDF, CDF
- Mean, Variance

Sample Space & Events

- **Random Experiment**: a repeatable procedure
- **Sample space**: set of all possible outcomes Ω.
- **Event**: a subset of the sample space.
- **Probability function**, P(ω): gives the probability for each outcome $\omega \in \Omega$
  - Probability is between 0 and 1
  - Total probability of all possible outcomes is 1.
  - If A=$\{\omega_1, \omega_2, \omega_3, \dots\}$, P(A)=P($\omega_1$)+P($\omega_2$)+P($\omega_3$) $+ \cdots$

Sample Space

- Discrete or continuous: countable (listable) or not?
- A sample space is discrete if it consists of a finite or countable infinite set of outcomes.
- A sample space is continuous if it contains an interval (or a union of multiple intervals) of real numbers.

Sample space - example

- Consider the random experiment in which items are selected from a batch consisting of three items {a,b,c}

- Case 1: select two items **without replacement**

  Sample space $\{ab, ac, ba, bc, ca, cb\}$

- Case 2: select two items **with replacement**

  Sample space $\{aa, ab, ac, ba, bb, bc, ca, cb, cc\}$

Events

- Events are sets:
  - ✓ Can describe in words
  - ✓ Can describe in notation
- Experiment: toss a coin 2 times.
- Event -- You get 1 or more heads

  = {HH, HT, TH}

Relation between two events

- Independence: Two events A and B are independent if and only if $P(A \cap B) = P(A) \cdot P(B)$

- Mutually Exclusive: Two events A and B are mutually exclusive if and only if $A \cap B = \emptyset$, which implies $P(A \cap B) = 0$

Random Variable

- X is called a random variable as it takes a numerical value that depends on the outcome of an experiment.

- Range of X: the set of possible values for X.

Probability Mass Function

- For a discrete random variable X with possible values $x_1, x_2, \dots, x_n$. A probability mass function $f(\cdot)$ is a function such that:

✓ $f(x_i) \geq 0$ for all $x_1, x_2, \dots, x_n$.

✓ $\sum_{i=1}^{n} f(x_i) = 1$

✓ $f(x_i) = \boxed{P(X = x_i)}$ for all $x_1, x_2, \dots, x_n$.

Probability that the random variable takes value x_i.

Probability Density Function

- For a continuous RV X, $P(X = x) \ = \ 0$ but $\{x\}$ is not an impossible event.

> - If $P(E) \ = \ 0$, then E is a zero-probability event.
> - If E is empty, then E is impossible.

- For a continuous random variable, we introduce a function f(ω), called the probability density function (pdf).
  - f(ω) > 0, if $\omega \in S$
  - f(ω) = 0, if $\omega \notin S$
  - $\int_{-\infty}^{\infty} f(x)dx = 1.$

Cumulative Distribution Function

- The cumulative distribution function (cdf) gives the probability that the random variable X is less than or equal to x and is usually denoted by $F(x)$

- $F(x) = P(X \leq x)$

- $f(x) = F(x) - \lim_{y \uparrow x} F(y)$

Mean and Variance

- Mean

$$E[X] = \Sigma_x x\, P(X = x) = \Sigma_x x\, f(x)$$

- Variance

$$Var[X] = \Sigma_x (x - \mathrm{E}[X])^2\, f(x)$$

Linearity: $E[\sum_i C_i X_i] = \sum_i C_i E[X_i]$

C_i is a constant

No assumption on X_i

Expectation of a function of X:

$$\mathrm{E}[g(X)] = \sum_x g(x) P(X = x) = \sum_x g(x) f(x)$$

Expectation of a constant: $E[C] = C$

Conditional Probability

- Given the realization of event A, the probability of event B may change

- (Conditional probability) If A and B are events with P(B)>0, then the conditional probability of A given B, denoted by P(A|B), is defined as $P(A|B) = \frac{P(A \cap B)}{P(B)}$

- (Independence) Two events A and B are called independent if and only if $P(A \cap B) = \mathrm{P}(A)P(B)$ $\quad P(A|B) = \frac{P(A \cap B)}{P(B)} = \mathrm{P(A)}$

Example: Conditional Probability in NLP

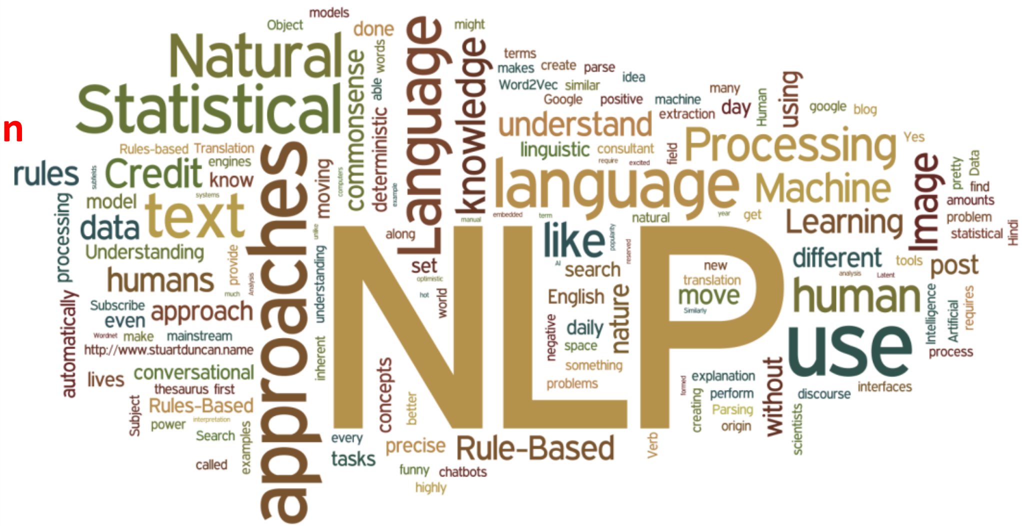
- H: mention “Interesting” in message
- D: mention “DDA2001” in message

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- P(H) = .1
- P(D) =.1
- P(H ∩ D) = .08

Known Information from data

- P(H|D) = 0.08/0.1 = 0.8



Discrete RV

- **Bernoulli distribution**
- **Binomial distribution**
- **Geometric distribution**

Bernoulli distribution

- Take value 1 with probability p and value 0 with probability 1 – p.

- $X \sim Bernoulli(p)$

- Mean=p
- Variance=p(1-p)

Binomial distribution

- N independent experiments
- Each experiment: success (with probability p) or failure (with probability 1 − p).
- X: the number of success (failure).

- $X \sim Binomial(N, p)$

- Mean=Np
- Variance=Np(1-p)

Geometric distribution

- Continuously draw a Bernoulli R.V.
- The X-th sample is the first success.
- X follows a geometric distribution.

- $X \sim Geometric(p)$ or $X \sim Geo(p)$

- Mean=1/p
- Variance=(1-p)/p^2

Continuous RV

- **Uniform Distribution**
- **Normal Distribution**

Uniform Distribution

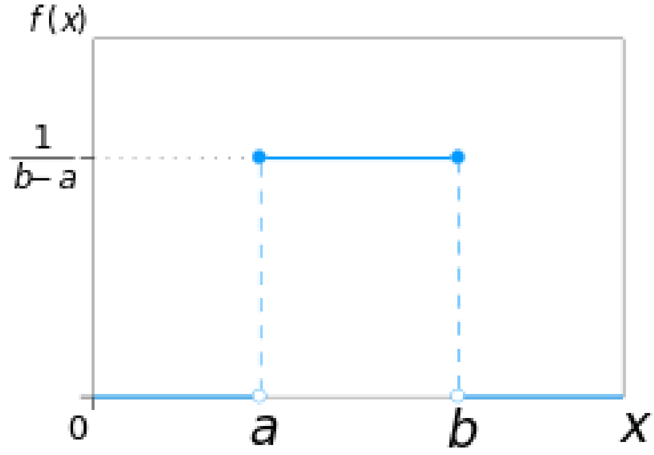
- With the same probability, X takes a value within [a, b], where b>a.

- f(x) = c for $x \in [a, b]$ and f(x) = 0 for $x \notin [a, b]$
- As $\int_{-\infty}^{\infty} f(x)dx = c(b - a) = 1,$ we have

$$c = \frac{1}{b - a}$$

Uniform Distribution

- With the same probability, X takes a value within [a, b]
- $X \sim Uniform(a, b)$
- Mean= $(a + b)/2$
- Variance=$(b - a)^2/12$



Applications

How to approximate $\int_0^2 e^{x^2+\cos(x)}dx$?

Applications

- Given $X \sim Uniform(0,2)$
- What's the value of $E[2\ e^{X^2+\cos(X)}]$?

- f(x) = ½ for $x \in [0,2]$
- $E[2\ e^{X^2+\cos(X)}] = \int_0^2 2\ e^{x^2+\cos(x)} f(x) dx = \int_0^2\ e^{x^2+\cos(x)} dx$

Given $X \sim Uniform(0,2)$, $E[2\, e^{X^2+\cos(X)}] = \int_0^2 e^{x^2+\cos(x)} dx$

- Draw N samples of $X \sim Uniform(0,2)$: $X_1, X_2, X_3, \ldots, X_N$
- Calculate $\frac{\sum_i 2\, e^{X_i^2+\cos(X_i)}}{N}$

Why? Expectation can be approximated by long-run average.

General Case

- How to calculate $\int_a^b h(x)dx$?

  - Draw N samples of $X \sim Uniform(a, b)$: X_1, X_2, X_3, …. , X_N
  - Calculate $\frac{\Sigma_i (\boldsymbol{b-a})\, h(X_i)}{N}$

- $E[h(x)]$ only gives you the average "height" of $h(x)$
- In order to get $\int_a^b h(x)dx$, which is the area, we need to multiply $E[h(x)]$ by $(b-a)$

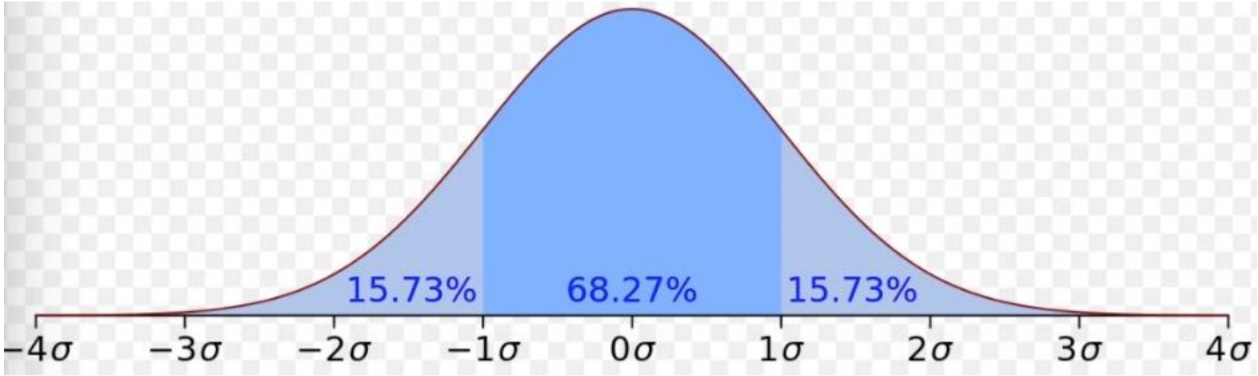
- Let $X \sim Uniform(a, b)$
- f(x) = 1/(b-a) for $x \in [a, b]$
- $E[(b-a)h(x)] = \int_a^b (b-a)h(x)f(x)dx = \int_a^b h(x)dx$

Normal Distribution

- X can be any real number
- Parameters: μ $(mean) and$ σ^2 $(variance)$

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

- $X \sim Normal(\mu, \sigma^2)$



Statistics

- Maximum Likelihood Estimation
- Linear Regression

Maximum likelihood estimate (MLE)

- Target: estimate θ of a model

- Samples: $X_1,\ X_2 \ldots,\ X_n$
- Possible models: $\theta \in \Theta$ (Θ depends on the information you have)
- Model performance (probability): L(θ)= P ($X_1,\ X_2 \ldots,\ X_n | \theta$)

- Estimator: $\hat{\theta}$ such that L(θ) is maximized at θ= $\hat{\theta}$.

Likelihood function

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

<u>Continuous RV model:</u>

- Likelihood: L(θ)= $\boxed{\Pi_i}$ f($X_i\,|\,\theta$)
- Log-Likelihood: $l(\theta)$= $\sum_i \log(\mathrm{f}(X_i\,|\,\theta))$

<u>Discrete RV model:</u>

- Likelihood: L(θ)= $\boxed{\Pi_i}$P ($X_i\,|\,\theta$)
- Log-Likelihood: $l(\theta)$= $\sum_i \log(\mathrm{P}(X_i\,|\,\theta))$

Why do we multiply the pdfs or pmfs here?

Because we assume those samples are independent observations

Example

- Suppose heights of people follow a normal distribution.
  - Given parameter μ, the model is $N(\mu, 1)$

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

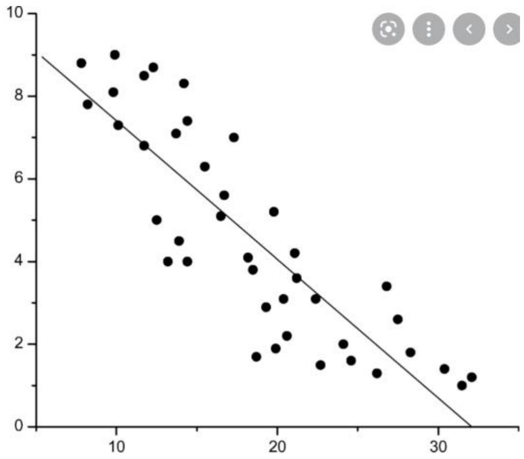
- L(μ) = $\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_i (X_i - \mu)^2}$
- $l(\mu)$ = Log L(μ) = $-\frac{1}{2}\sum_i (X_i - \mu)^2$ - $\frac{\mathrm{n}}{2}\log 2\pi$

- $l'(\mu)$ = $\sum_i (X_i - \mu)$= 0 $\Longrightarrow$ $\hat{\mu} = \bar{X}.$

Linear Regression

- Step 1: Propose a model ($Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$)
- Step 2: Estimate β_0, β_1 (Maximum Likelihood Estimate)
- Step 3: Check assumptions (residual analysis)

Propose a model



- $Y \sim N\left(\beta_0 + \beta_1 X, \sigma^2\right)$

- Given the observation of X, Y follows a normal distribution with mean $\beta_0 + \beta_1 X$, and variance σ^2

- To simplify the analysis, we assume σ^2 is known

Estimate β_0, β_1

- Samples: $(X_1, Y_1), \ldots, (X_N, Y_N)$
- For the model with β_0,β_1, σ^2 , the likelihood function is

$$\frac{1}{(\sqrt{2\pi})^n \sigma^n} \exp\left[-\frac{1}{2} \frac{\Sigma_i (Y_i - \beta_1 X_i - \beta_0)^2}{\sigma^2}\right]$$

- Given σ^2, to maximize the likelihood, we only need to minimize

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0)^2$$

- Taking derivative over β_0 and β_1, we have

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0) = 0$$

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0) X_i = 0$$

Estimate β_0, β_1

$$\Sigma_i(Y_i - \beta_1 X_i - \beta_0)X_i = 0 \quad \text{AND} \quad \Sigma_i(Y_i - \beta_1 X_i - \beta_0) = 0$$

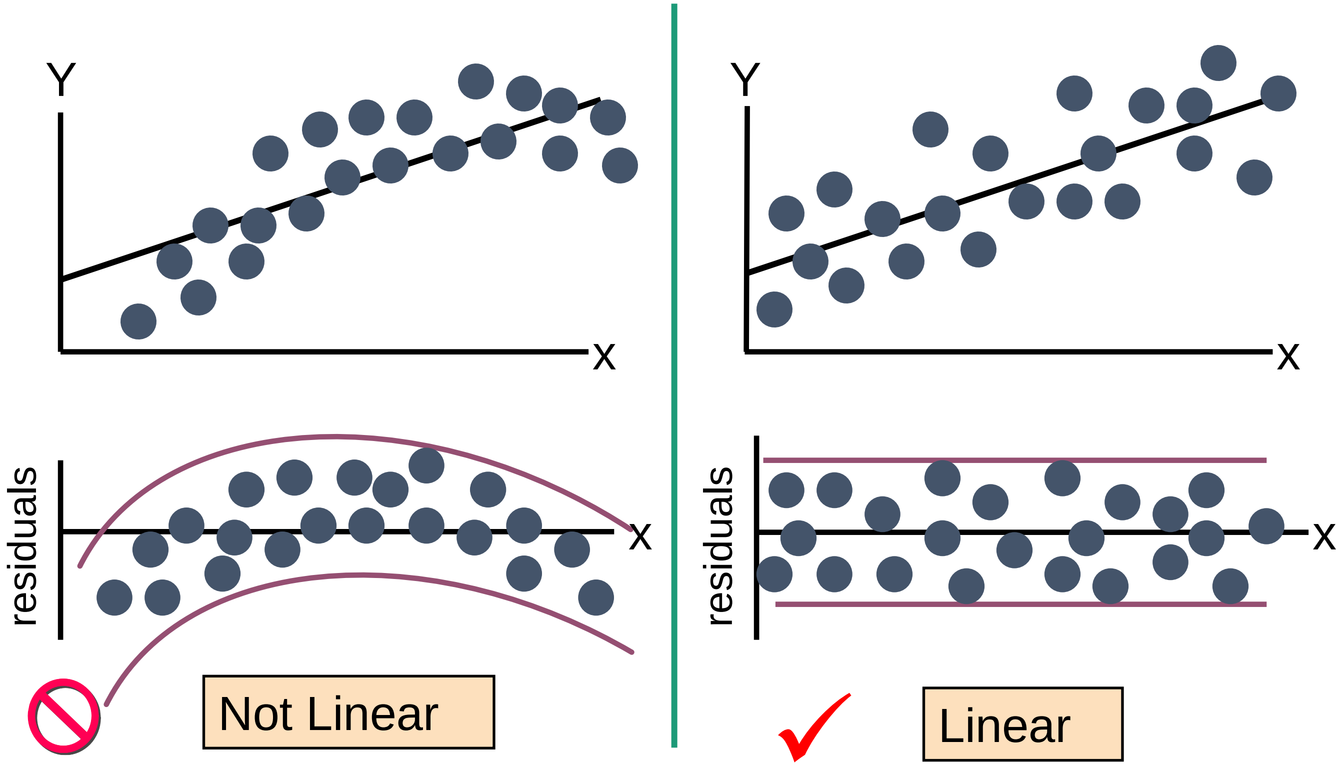
Eliminate β_0 first:

$$\Sigma_i(Y_i - \beta_1 X_i - \beta_0) = 0 \quad \text{->} \quad \beta_0 = \frac{1}{N}\Sigma_i(Y_i - \beta_1 X_i) = \bar{Y} - \beta_1 \bar{X}$$

MLE:

$$\begin{cases} \widehat{\beta_1} = \dfrac{\Sigma_i(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma_i(X_i - \bar{X})^2} \\ \widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X} \end{cases}$$

Residual Analysis for Linearity



Residual Analysis for constant-variance