Probability

CS 155 Machine Learning and Data Mining Recitation

Lucy Yin
January 21, 2016

Motivations

- **Uncertainty** is everywhere around us
  - "what is the chance of raining today?"
  - "when will the next bus arrive?"
  - "will I go to the recitation today?"
- Machine learning tries to understand uncertainties and interact with the real world
- **Probability theory** is the mathematical study of uncertainty.

Basic Concepts

- Sample Space Ω: set of all possible outcomes
- Event A is a subspace of Ω
  - P(A) ≥ 0 (non-negativity)
  - P(Ω) = 1 (trivial event)
  - For 2 events A and B: (addictivity)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Basic Concepts

- **Example**: rolling a fair 6-sided dice
  - $\Omega=\{1,2,3,4,5,6\}$
  - $P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = 1/6$
  - $P(\{2,4,6\}) = P(\{2\})+P(\{4\})+P(\{6\}) = 1/2$

Joint and Conditional Probability

For a pair of events x and y:

- **Joint Probability** is the probability of both events occurring at the same time: P(x,y)

$$0 \leq P(x,y) \leq 1$$

$$\sum_x \sum_y P(x,y) \leq 1$$

(Discrete RV)

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x,y) dxdy \leq 1$$

(Continuous RV)

- **Conditional Probability** x|y is the probability of event x if we consider only the cases in which y occurs: P(x|y)

Conditional Probability

Joint Probability

$$P(x|y) = \frac{P(x,y)}{P(y)} \qquad P(y) \neq 0$$

Joint and Conditional Probability

Example: Draw 2 Kings from a Deck

Event A=drawing a King first

Event B=drawing a King second

For the first card, the chance of drawing a King is 4/52 (there are 4 Kings in a deck of 52 cards)

$$P(A) = 4/52$$

After removing a King from the deck, the probability of the 2nd card drawn is less (only 3 Kings left in the remaining deck)

$$P(B|A) = 3/51$$

And so:

$$P(A,B) = P(B|A)P(A) = \frac{3}{51} * \frac{4}{52} = \frac{12}{2652} = \frac{1}{221} \approx 0.5\%$$

So, the chance of getting a pair of Kings is about 0.5%

Marginal Distribution

- If X and Y have a joint distribution with probability function p(x,y), then the **marginal distribution of X** has a probability function p(x), which is defined as

$$p(x) = \sum_{y} p(x,y)$$

(Discrete RV)

$$p(x) = \int_{-\infty}^{\infty} p(x,y)dy$$

(Continuous RV)

- Similarly, the marginal distribution of y is

$$p(y) = \sum_{x} p(x,y)$$

(Discrete RV)

$$p(y) = \int_{-\infty}^{\infty} p(x,y)dx$$

(Continuous RV)

Marginal Distribution

- Example:

	x_1	x_2	x_3	x_4	$p_y(Y)\downarrow$
y_1	$4/32$	$2/32$	$1/32$	$1/32$	$8/32$
y_2	$2/32$	$4/32$	$1/32$	$1/32$	$8/32$
y_3	$2/32$	$2/32$	$2/32$	$2/32$	$8/32$
y_4	$8/32$	0	0	0	$8/32$
$p_x(X) \rightarrow$	$16/32$	$8/32$	$4/32$	$4/32$	$32/32$

$$p(x_1) = \sum_y p(x_1, y) = p(x_1, y_1) + p(x_1, y_2) + p(x_1, y_3) + p(x_1, y_4)$$

$$= {}^4/_{32} + {}^2/_{32} + {}^2/_{32} + {}^8/_{32} = {}^{16}/_{32}$$

Independence

- Event A, B are independent:

$$P(A,B) = P(A)P(B)$$

or equivalently

$$P(A|B) = P(A)$$

Recall

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

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

$$P(A|B) = P(A)$$

Independence

Example:

Roll a dice twice. What is the probability of rolling 6 at both trials?

A=rolling a 6 in the first trial

B=rolling a 6 in the second trial

$$P(A,B) = P(A)P(B)$$
$$= {}^{1}/_{6} * {}^{1}/_{6} = {}^{1}/_{36}$$

Joint Probability Distribution

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

- Chain Rule

$$P(A_1, A_2, \ldots, A_n) = P(A_n, \ldots, A_2, A_1)$$

$$= P(A_n|A_{n-1} \ldots, A_2, A_1)P(A_{n-1} \ldots, A_2, A_1)$$

$$\ldots$$

$$= P(A_n|A_{n-1} \ldots, A_2, A_1)P(A_{n-1}|A_{n-2} \ldots, A_2, A_1) * \ldots * P(A_2|A_1)P(A_1)$$

$$= \prod_{i=1}^{n} P(A_i|A_1,\ A_2\ , \ldots, A_{i-1})$$

Bayes' Theorem

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

Likelihood Function

Prior Information

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

Posterior Probability

Evidence

$$P(A|B) \propto P(B|A)P(A)$$

Bayes' Theorem

Example: If a person has an allergy (A), very often sneezing (S) is observed

$$P(S|A) = 0.8$$

What is the chance of an allergy is sneezing is observed?

$$P(A|S) = ?$$

More information: P(A) = 0.001 (assume very little people has allergy), P(S) = 0.1 (assume many people sneeze)

$$P(A|S) = \frac{P(S|A)P(A)}{P(S)}$$

$$= \frac{0.8 * 0.001}{0.1} = 0.008$$

So, 0.8% chance the sneezing is due to allergy.

Random variable

- Random variable X is a function X: Ω -> R
  - Example: number of heads in 20 tosses of a coin
  - Discrete and continuous random variable
- Cumulative Distribution Function (CDF):

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

  - Properties:
    - $0 \leq F(x) \leq 1$
    - $F(x)$ is monotonically increasing
    - $\lim_{x \to -\infty} F(x) = 0 \qquad \lim_{x \to +\infty} F(x) = 1$

Discrete random variable

- r.v. of the underlying distribution can take only *finite* many different values
- Probability Mass Function (pmf):

$$p(x) = P(X = x)$$

  – Example:
    - Rolling a dice

X	1	2	3	4	5	6
P(X)	1/6	1/6	1/6	1/6	1/6	1/6

Continuous random variable

- r.v. of the underlying distribution can take *infinite* many different values
- Probability Density Function (pdf)

$$f(x) = \frac{dF(x)}{dx}$$

  – Knowing cdf, we can calculate $P(a < x \leq b)$ for all intervals from a to b

Expectation

- Expectation: mean of the distribution
- Expectation for random variables X: $E(x)$
  - Discrete X: $E(x) = \sum_x xp(x)$
  - Continuous X: $E(x) = \int_x xf(x)$
- Expectation is linear

$$E(aX) = aE(X) \qquad a \text{ is const}$$
$$E(X+Y) = E(X) + E(Y)$$

Variance

- Variance of a distribution is the measure of the “spread” of a distribution.

$$Var(X) = E\left(\left(X - E(X)\right)^2\right)$$

or equivalently

$$Var(X) = E(X^2) - E(X)^2$$

- Variance is NOT linear

$$Var(aX + b) = a^2 Var(X) \qquad a, b \text{ is const}$$

Some Important Distributions

- Bernoulli(p)

$$p(x) = p^x(1-p)^{1-x} \quad for\ x = 0,1 \quad E(x) = p$$

- Binomial(n,p)

$$p(x) = \binom{n}{x} p^x(1-p)^{n-x} \quad E(x) = np$$

- Geometric(p)

$$p(x) = p(1-p)^{x-1} \quad E(x) = {}^1\!/_p$$

- Poisson(λ)

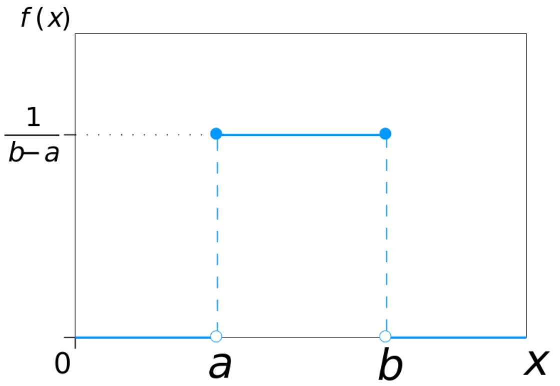
$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad E(x) = \lambda$$

Some Important Distributions

- Uniform (a,b) (a<b)

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & otherwise \end{cases}$$

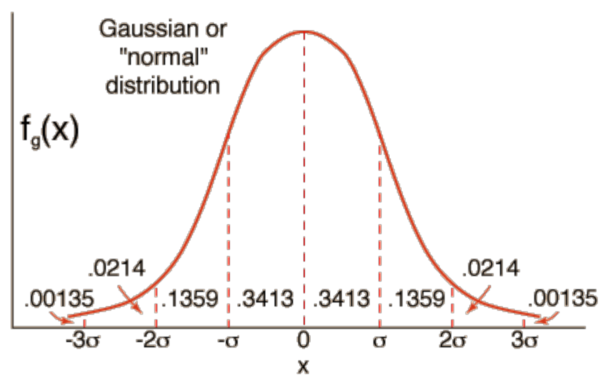
$$E(x) = \frac{1}{2}(a+b)$$



- Normal (μ, σ^2)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$E(x) = \mu$$



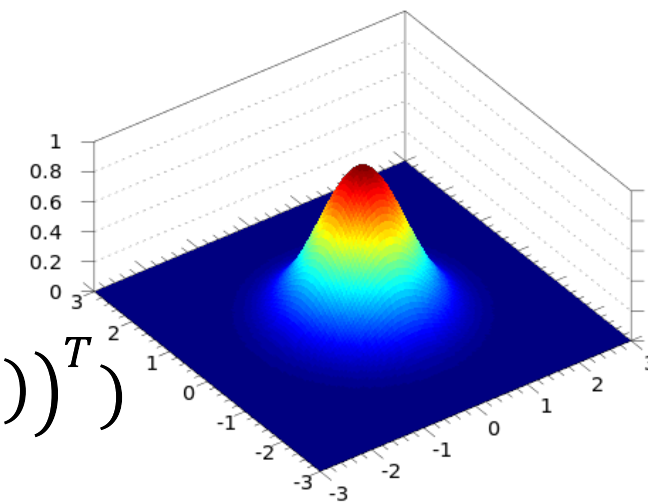
Multivariate Gaussian Distribution

- $X = [X_1, X_2, \ldots, X_n]^T$ random vector
- $X \sim \mathcal{N}(\mu, \Sigma)$ n-dimensional Gaussian distribution:

$$f(X) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(X-\mu)^T \Sigma^{-1} (X-\mu)\right)$$

$$E(x) = \mu$$

$$Cov(x) = \Sigma$$
$$= E((X - E(X)\big(X - E(X)\big)^T)$$



Example of a 2D Gaussian Distribution

Parameter Estimation

- Parametrized distribution P(x,θ) with parameter(s) θ unknown
- iid samples $x_1, x_2, \ldots, x_n$ observed
- Goal: estimate θ
- Recall Bayes' Theorem: $P(\theta|X) \propto P(X|\theta)P(\theta)$
  - (ideally) MAP: $\hat{\theta} = argmax\, P(\theta|X)$
  - (in practice) MLE: $\hat{\theta} = argmax\, P(X|\theta)$

Parameter Estimation – "log" trick

- Logarithmic function is monotonically increasing, it will not distort where the maximum is location (although the maximum value of the function before and after taking logarithm will be different)
- Simplify the calculation
  - Gradient descent could be used for minimization
  - Multiplication turns into summation

$$argmax_\theta \, f(\theta|x) = argmin_\theta - \log\left(f(\theta|x)\right)$$

Parameter Estimation

- **Example 1**: Binomial distribution
- Coin toss. Repeat the tossing experiment n times, and observe k time 'head'
- What is the probability observing head?

$$\text{argmax}_p\, P(k|p) = argmax \binom{n}{k} p^k (1-p)^{n-k}$$

Parameter Estimation

- Example 1: Binomial distribution

$$\text{argmax}_p\, P(k|p) = argmax \binom{n}{k} p^k (1-p)^{n-k}$$

$$= argmax \;\; p^k (1-p)^{n-k}$$

$$= argmin - \log\left(p^k (1-p)^{n-k}\right)$$

$$= argmin - k\log p - (n-k)\log(1-p)$$

Take derivatives wrt p and zeroing: $p = \dfrac{k}{n}$

Parameter Estimation

- **Example 2**: Gaussian distribution
- Give $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$ data samples, what is the optimal μ and σ^2 assuming independence of the observed data

$$\mathrm{argmax}_{\mu,\sigma^2} P\left(x^{(1)}, \ldots, x^{(n)} | \mu, \sigma^2\right)$$

$$= \mathrm{argmax}_{\mu,\sigma^2} \left(\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}\left(x^{(1)}-\mu\right)^2}\right) \ldots \left(\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}\left(x^{(n)}-\mu\right)^2}\right)$$

$$= \mathrm{argmax}_{\mu,\sigma^2} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}\left(x^{(i)}-\mu\right)^2}$$

Parameter Estimation

- Example 2: Gaussian distribution

$$\text{argmax}_{\mu,\sigma^2} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}\left(x^{(i)}-\mu\right)^2}$$

$$= \text{argmin}_{\mu,\sigma^2} - log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}\left(x^{(i)}-\mu\right)^2}\right)$$

$$= \text{argmin}_{\mu,\sigma^2} - \sum_{i=1}^{n}\left(log\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2}\frac{(x^{(i)}-\mu)^2}{\sigma^2}\right)$$

$$= \text{argmin}_{\mu,\sigma^2} \quad \frac{n}{2}\log(\sigma^2) + \frac{n}{2}\log(2) + \frac{1}{\sigma^2}\sum_{i=1}^{n}\left((x^{(i)}-\mu)^2\right)$$

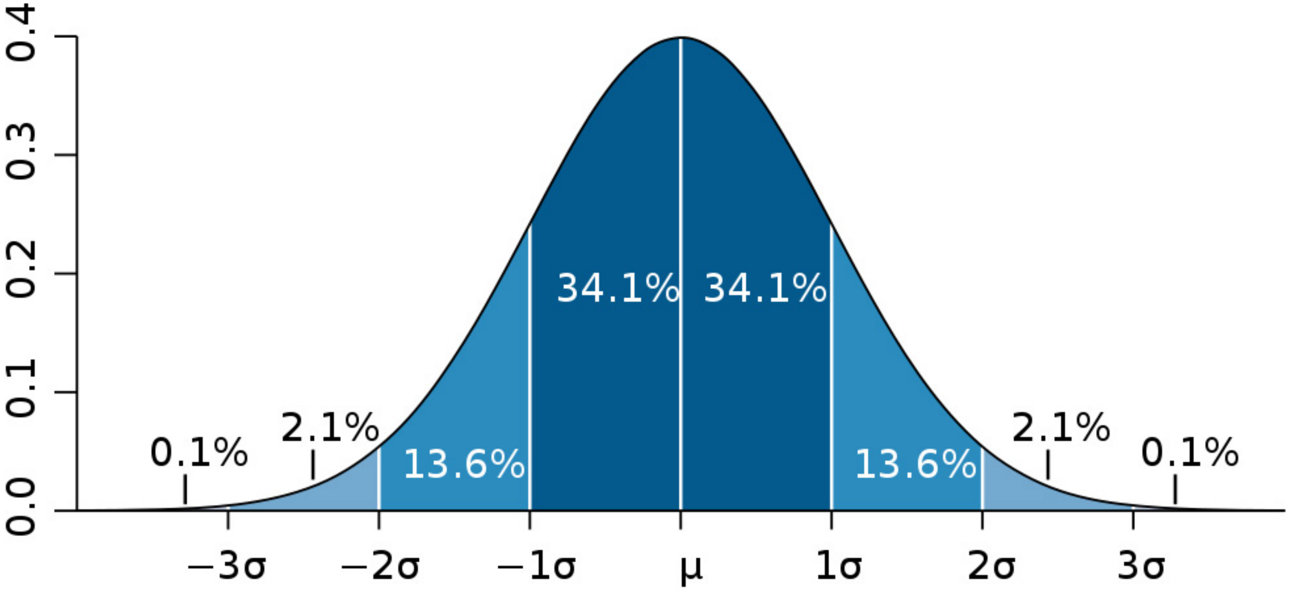
Take partial derivatives wrt μ and σ^2 and zeroing…

Central Limit Theorem

- Central limit theorem: Let $X_1, X_2, \ldots, X_n$ be iid with finite mean μ and finite variance σ^2, then the random variable $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately Gaussian with mean μ and variance σ^2/n
- Approximation becomes better as n grows

Confidence Interval

- A **Confidence interval** is an interval in which a measurement or trial falls corresponding to a given probability.



$$P(\mu - \sigma \leq x \leq \mu + \sigma) \approx 0.6827$$
$$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 0.9545$$
$$P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 0.9973$$

Hypothesis testing

- **Null Hypothesis (H_0)**: A maintained hypothesis that is held to be true unless sufficient evident to the contrary is presented.
- **Alternative Hypothesis (H_1)**: A hypothesis that is held to be true when the null hypothesis is rejected.
- **Significance Level (α)**: The probability of rejecting a true null hypothesis.
- **P-value**: The probability of obtaining the observed sample results assuming the null hypothesis is actually true
- Decision Criterion for a Hypothesis Test using P-value:
  - p-value $< \alpha$ => reject H_0
  - P-value $> \alpha$ => fail to reject H_0

Hypothesis testing

- Example: IQ is normally distributed in the population according to a $N(100, 15^2)$ distribution. We tested 9 Caltech students and find they have an average IQ of 112.

  H_0: Caltech students' IQ follow a $N(100,15^2)$ distribution

  H_1: the average Caltech student IQ is greater than 100

- Can we reject H_0 at a significant level α =0.05?
- z-statistic

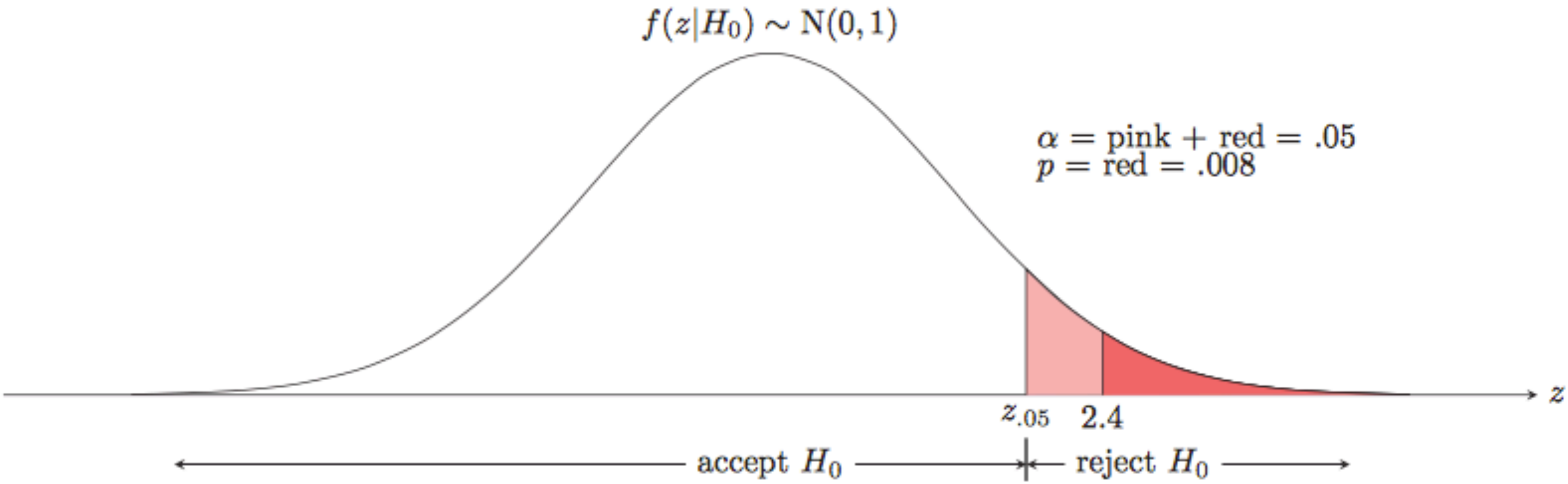
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{112 - 100}{15/\sqrt{9}} = 2.4$$

$$p = P(z \geq 2.4) = 0.0081975$$

$$p < \alpha$$

Hypothesis testing

- Can we reject H_0 at a significant level α =0.05?



Reject H_0: in favor of the alternative hypothesis that Caltech students have higher IQ than average