CSC380: Principles of Data Science

Midterm review

Xinchen Yu

Acknowledgement: Built on Jason Pacheco, Kwang-Sung Jun, Chicheng Zhang's slides

Midterm Info

- What you can bring?
  - Cheat sheet: letter size, double-sided
  - Scientific calculator

$\frac{\sqrt{2}}{3}$ is ok

- Time: Oct 12, Thursday, 3:30-4:45pm

- Location: C E Chavez Bldg, Rm 111 (same as lecture room)

General tips on midterm preparation

- Prioritize reviewing basic concepts & ideas
- Understand the motivations and links between concepts
- “Memorization with understanding”

- Try to solve these on your own, then discuss with classmates
  - examples in the slides
  - practice problems
  - HW questions (esp. if you did not get them right the first time)

General tips on midterm preparation

- What will not included in the midterm?
  - Code related questions
  - Data analysis and visualization
  - Pure proof questions
    - may need you to provide justifications

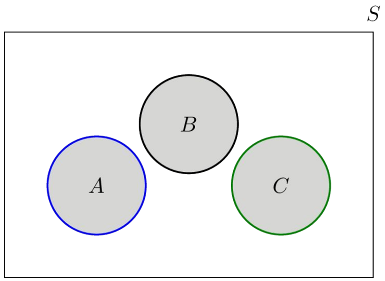
Office hours

- Office hours only for midterm:
  - Hui Ni: Oct 6, this Friday 2:00-3:30pm, GS 856
  - Saiful: Oct 9, next Monday 12:30-2:00pm, GS 934
  - Xinchen: Oct 10, next Tuesday 12:30-2:30pm, GS 854
  - Shahriar: Oct 11, next Wednesday 10:00-11:30am, zoom

Probability

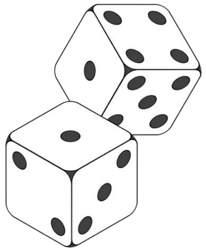
Probability

- Basic definitions: outcome space, events



- Probability P: maps events to [0, 1] values
  - Three axioms
  - Axiom 3: additivity

- Special case of P: each outcomes is equally likely

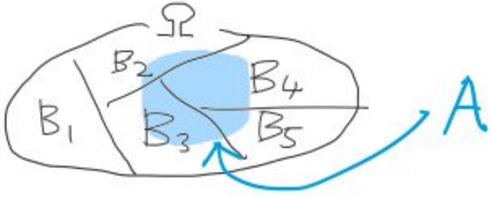


$$P(E) = \frac{|E|}{|\Omega|}$$

Number of elements in event set

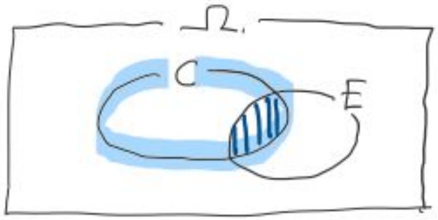
Number of possible outcomes (36)

- distributive law, inclusion-exclusion rule; law of total probability



Probability

- $P(E \cap C) = P(E|C)P(C) = P(C|E)P(E)$



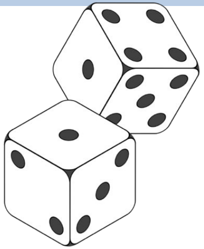
- Conditional probability
  - Chain rule, chain rule + law of total probability, bayes rule
  - Important application: medical diagnosis
  - Approach: write down the joint probability table

- Independence of events:

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

- Conditional / joint / marginal probability

Probability



- Discrete random variable X (e.g., sum of two dice)

- Representation of its distribution: probability mass function (PMF)
  - Tabular representation of joint distribution of 2 RVs (X,Y)

- RVs: law of total probability, conditional probability, chain rule, bayes rule, independence, conditional independence

- Useful discrete distributions
  - Uniform
  - Bernoulli
  - Binominal

Probability

- Continuous random variable X: $P(X = x) = 0$ for any x

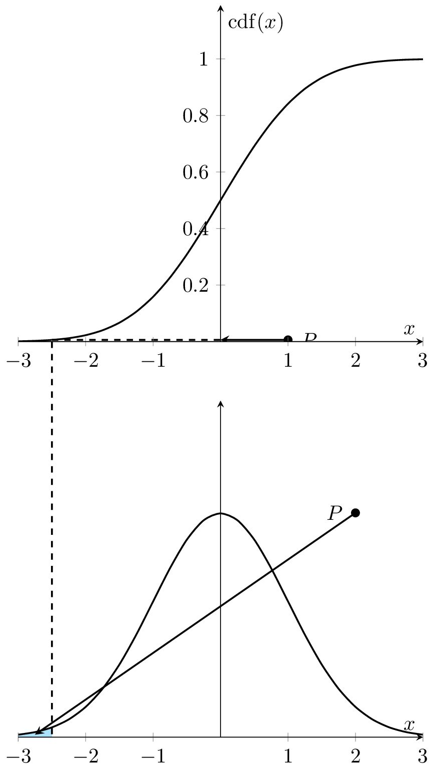
- Probability density function (PDF)

$$P(a < X \leq b) = \int_a^b p(x)\mathrm{d}x \quad p(x) = \frac{dF(x)}{dx}$$

- Cumulative distribution function (CDF)

$$P(a < X \leq b) = F(b) - F(a)$$

- Useful continuous distributions
  - Uniform
  - Gaussian (important properties)



Probability

- Moments of random variables: expectation, variance, covariance

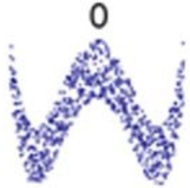
- Calculate mean (expectation) and variance of RVs
    - Linearity of expectation: $E[X + cY] = E[X] + cE[Y]$ for constant c
    - $E[X^2]$
    - $E[XY]$
        - If independent: $E[X]E[Y]$
        - If not independent: $E[XY] = \sum_{(x,y)} \sum xy \cdot p(x, y)$
    - $E[X \mid Y = y]$
    - $Var[cX]$
    - $Var[X + c] = Var[X]$ (not in slides, prove?)
    - $Var[X+Y]$ when independent

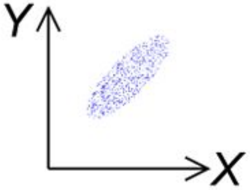
$$\boxed{\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2]}$$

- Expectation and variance of useful distributions
    - Bernoulli
    - Gaussian

- Measures *linear relationship* between X, Y

$$\mathrm{Cov}(X,Y) = 0 \not\Rightarrow X \perp Y$$





- Pearson correlation: $\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$, where $\sigma_X = \sqrt{\mathrm{Var}(X)}$

- Important property: $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$
    - What if X, Y are independent?

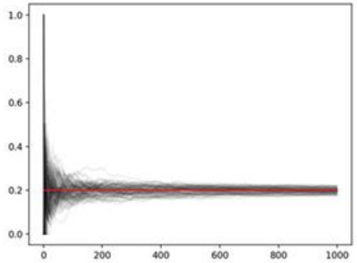
Statistics

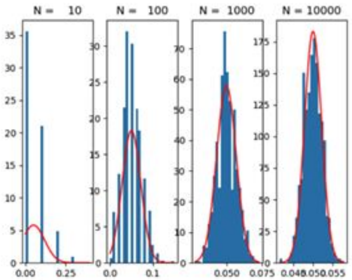
Statistics

- Statistics: make statements about data generation process based on data seen; reverse engineering

- Point estimation
  - Given iid samples $X_1, \dots, X_n \sim \mathcal{D}_\theta$, estimate θ by constructing *statistics* $\hat{\theta}_n$
  - Basic estimators: sample mean, sample variance
  - Performance measures: unbiasedness, consistency, MSE (efficiency)
  - Bias-variance decomposition:
    - $\mathrm{MSE}(\hat{\theta}) = \mathrm{bias}(\hat{\theta})^2 + \mathrm{var}(\hat{\theta})$

- Useful probability tools:
  - Law of Large Numbers
  - Central Limit Theorem





Statistics

- Sample mean, sample variance

- Sample variance
  - biased version
  - unbiased version
  - how to determine an estimator is biased or unbiased?

- MSE, Bias, Variance
  - how to calculate expectation and variance if there are more than 1 random variable -- use what we learned in probability lecture 5 & 6

Statistics

- Calculate bias and variance

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}_n) &= \mathbf{E}[(\hat{\theta}_n - \theta)^2] \\ &= \Big(\mathbf{E}[\hat{\theta}] - \theta\Big)^2 + \mathbf{E}[(\hat{\theta} - \mathbf{E}[\hat{\theta}])^2] \\ &= \mathrm{bias}^2(\hat{\theta}) + \mathrm{Var}(\hat{\theta})\end{aligned}$$

Important properties of Gaussian

- Closed under additivity:

$$X \sim \mathcal{N}(\mu_x, \sigma_x^2) \qquad Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$$

$$X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

- Closed under affine transformation (a and b constant):

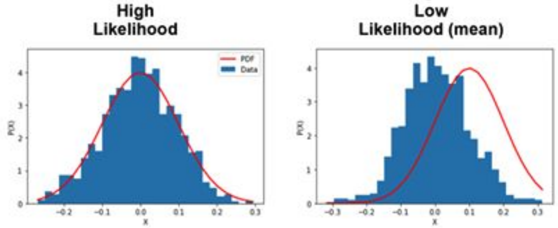
$$aX + b \sim \mathcal{N}(a\mu_x + b, a^2\sigma_x^2)$$

Statistics

- Maximum likelihood (MLE): a general approach for point estimation

- Given $X_1, \ldots, X_n \sim \mathcal{D}_{\theta^*}$, estimate θ^* by finding the maximizer of the likelihood function

$$\mathcal{L}_n(\theta) = p(x_1, \ldots, x_n; \theta) = p(x_1; \theta) \cdot \ldots \cdot p(x_n; \theta)$$



- Intuition: $\mathcal{L}_n(\theta)$ measures the "goodness of fit" of $\mathcal{D}_\theta$ to data $x_1, \ldots, x_n$

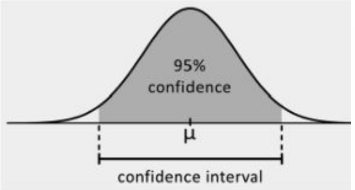
- $\mathcal{D}_\theta$ can be general, e.g. Bernoulli, Gaussian, Poisson (in HW3)

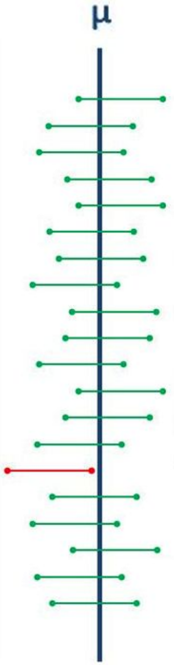
Statistics

- Confidence interval (interval estimation)

- Definition of confidence intervals:
  - Given data $X_1, \dots, X_n \sim \mathcal{D}_\theta$ with unknown θ (say, $\mathcal{D}_\theta = \mathcal{N}(\theta, 1)$)
  - Construct a_n, b_n (that depends on $X_1, \dots, X_n$), such that

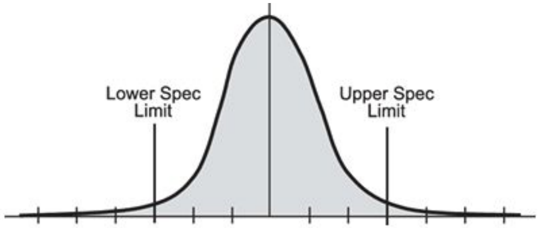
$$P(\theta \in [a_n, b_n]) \geq 1 - \alpha$$





- Interpretation: unless we are extremely unlucky (in that we encounter an unrepresentative dataset, which happens with prob. $\leq \alpha$), our confidence interval always contains the underlying parameter

Statistics



- Confidence intervals for population mean:
  - Gaussian(naive):
    $\left[\hat{\mu} - \frac{z_{1-\alpha/2}\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + \frac{z_{1-\alpha/2}\hat{\sigma}}{\sqrt{n}}\right]$, $z_{1-\alpha/2} = 1 - \alpha/2$-quantile of $\mathcal{N}(0,1)$
  - Gaussian(corrected):
    $\left[\hat{\mu} - \frac{t_{1-\alpha/2}\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + \frac{t_{1-\alpha/2}\hat{\sigma}}{\sqrt{n}}\right]$, $t_{1-\alpha/2} = 1 - \alpha/2$-quantile of t distribution (degree of freedom=?)

  - We expect you to be able to compute them on a small dataset

- Confidence intervals for general population parameters: bootstrap