CSE312 Final Notes

Counting:

- $P(A) = \frac{|A|}{|\Omega|}$, where A is desired event and Ω is the sample space.
- Product rule: if 1st step has m choices and 2nd step has n choices, 2 steps together have mn choices.
- Permutation (order matters): pick k objects from n and permutes

$$P(n,k) = \frac{n!}{(n-k)!}$$

- Combination (order doesn't matter): choose r objects from n / n choose r

$$\binom{n}{r} = \frac{P(n,r)}{r!} = \frac{n!}{r!\,(n-r)!}$$

  Identity: $\binom{n}{r} = \binom{n}{n-r} = \binom{n-1}{r-1} + \binom{n-1}{r}$

  Binomial Theorem: $(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}$ Corollary: $\sum_{i=0}^{n} \binom{n}{i} = 2^n$
- Complementing: P(contains at least 1) = 1 – P(contains 0)
- Inclusion-Exclusion: + single – pairs + triples – quads
- Pigeonhole Principle: If you have n pigeons and k holes, then some hole has > 1 pigeon.

Probability:

- 2 events E and F are *mutually exclusive* if and only if $E \cap F = \emptyset$
- Axioms of Probability:
  - $P(E) \geq 0$
  - $P(\Omega) = 1$
  - If E and F are mutually exclusive, then $P(E \cup F) = P(E) + P(F)$

  Implications of Axioms:
  - $P(\bar{E}) = 1 - P(E)$
  - If $E \subseteq F$, then $P(E) \leq P(F)$
  - $P(E) \leq 1$
  - $P(E \cup F) = P(E) + P(F) - P(E \cap F)$
- Equally likely outcomes: $P(a) = \frac{1}{|\Omega|}$ for every $a \in \Omega$
- Conditional Probability: suppose conditional probability of E given F

$$P(E|F) = \frac{|E \cap F|}{|F|} = \frac{P(E \cap F)}{P(F)}$$

  Chain rule: $P(E_1 \cap E_2 \cap \ldots \cap E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1E_2) \ldots P(E_n|E_1E_2 \ldots E_{n-1})$
- Law of Total Probability: $P(E) = P(E|F)P(F) + P(E|\bar{F})P(\bar{F})$
- Conditional Independence: A and B are conditionally independent if and only if

$$P(A \cap B|C) = P(A|C)P(B|C)$$

- Bayes' Theorem: $P(F|E) = \frac{P(E|F)P(F)}{P(E)}$ Corollary: $P(F|E) = \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|\bar{F})P(\bar{F})}$
- Naive Bayes Classifier:

$$P(Spam|x_1, x_2, \ldots, x_n) \approx \frac{P(Spam)\prod_{i=1}^{n} P(x_i|Spam)}{P(Spam)\prod_{i=1}^{n} P(x_i|Spam) + P(Ham)\prod_{i=1}^{n} P(x_i|Ham)}$$

  - Assumption: words in the email are conditionally independent given we know the email is spam/ham.
  - $P(Spam)$ and $P(Ham)$ are fractions of spam/ham emails in training data.
  - Laplace Smoothing for each $P(word|spam)$
- Independent Events: E and F are independent if and only if $P(E \cap F) = P(E)P(F)$

  If $P(F) > 0$, then E and F are independent if and only if $P(E|F) = P(E)$.

Discrete random variables and expectation:

- Random Variable: numerical function of the outcome
  (Discrete: countable number of possible values.)
- Independent Random Variables: Random Variables X and Y are independent if and only if

$$\forall x \forall y \; P(X = x \wedge Y = y) = P(X = x)P(Y = y)$$

- Probability Mass Function (pmf): Let T be the set of outcomes and a be an outcome:

$$p(a) = \begin{cases} P(X = a)\text{, if } a \in T \\ 0\text{, otherwise} \end{cases}$$

- Expectation of a random variable: $E[X] = \sum_x xp(x)$
  - Linearity of Expectation: $E[aX + b] = aE[X] + b$, $E[X + Y] = E[X] + E[Y]$
  - If X and Y are independent, $E[XY] = E[X]E[Y]$
  - Indicator Random Variable: $X_i = \begin{cases} 1 \\ 0 \end{cases}$ for $0 \le i \le n$
- Variance (σ^2) and Standard deviation (σ):
  - Let $E[X] = \mu$, then $Var(X) = E[X - \mu]$ $\quad Std(X) = \sqrt{Var(X)}$
  - Theorem: $Var(X) = E[X^2] - (E[X])^2$ $\quad Var(aX + b) = a^2 Var(X)$
  - If X and Y are independent, $Var(X + Y) = Var(X) + Var(Y)$
- Distributions:
  - Uniform: $X \sim Unif(a, b)$ if X is equally likely to be any integer in $[a, b]$.
    - $p(X) = \frac{1}{b-a+1}$
    - $E[X] = \frac{1}{2}(b + a)$ $\quad Var(X) = \frac{1}{12}(b - a)(b - a + 2)$
  - Bernoulli: $X \sim Ber(p)$ is a random indicator variable with $P(X = 1) = p$ and $P(X = 0) = 1 - p$.
    - $E[X] = p$ $\quad Var(X) = p(1 - p)$
  - Binomial: $X \sim Bin(n, p)$ is the sum of n independent Bernoulli random variables such that $X_i = Ber(p)$ for $1 \le i \le n$.
    - $P(X = i) = \binom{n}{i} p^i (1 - p)^{n-i}$
    - $E[X] = np$ $\quad Var(X) = np(1 - p)$
  - Geometric: $X \sim geo(p)$ is independent Bernoulli trials with parameter p until and including 1st success.
    - $p(X = k) = (1 - p)^{k-1} p$ for $k \in \{1, 2, \dots\}$
    - $E[X] = \frac{1}{p}$ $\quad Var(X) = \frac{1-p}{p^2}$
  - Poisson: $X \sim Poi(\lambda)$ when evets happen independently with average rate of λ per unit time.
    - $P(X = i) = e^{-\lambda} \frac{\lambda^i}{i!}$
    - $E[X] = \lambda$ $\quad Var(X) = \lambda$
- Summations:
  - $\sum_{i=0}^{\infty} r^i = \frac{1}{1-r}$
  - $\sum_{i=1}^{n} i = \frac{n(n+1)}{2} = \binom{n+1}{2}$
  - $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$

Continuous Random Variables

- Continuous Random Variables: takes values from an uncountable set.
- Probability Density Function: $f_X(x)$
- Cumulative Distribution Function: $P(X \le x) = F(x) = \int_{-\infty}^{x} f_X(t)\, dt$, thus $F_X'(x) = f_x(x)$
- Distributions:
  - Uniform: $X \sim Unif(a, b)$ indicates each real number from $[a, b]$ to be equally likely.
    - $f_X(x) = \frac{1}{b-a}, \quad x \in [a, b]$
    - $E[X] = \frac{a+b}{2}$ $\quad Var(X) = \frac{(b-a)^2}{12}$

- Exponential: $X \sim Exp(\lambda)$ represents the waiting time to the first success where $\lambda > 0$ is the average number of events per unit time.
    - $f_X(x) = \lambda e^{-\lambda x}, x \geq 0$
    - $E[X] = \frac{1}{\lambda}$ $\quad$ $Var(X) = \frac{1}{\lambda^2}$
    - $F_X(x) = 1 - e^{-\lambda x}, x \geq 0$
    - Memoryless: for any $s, t \geq 0$, $P(X > s + t | X > s) = P(X > t)$
- Normal: $X \sim N(\mu, \sigma^2)$ if X has the probability density function of

$$f_X(x) = \frac{1}{\sigma\sqrt{\{2\pi\}}}\ e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, x \in R$$

    - $E[X] = \mu$ $\quad$ $Var(X) = \sigma^2$
    - Standard normal: $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$
      $\Phi(z) = F_Z(z) = P(Z \leq z)$ $\quad$ $\Phi(-z) = 1 - \Phi(z)$
    - Closure of Normal Distribution: linear transformation of normal is still normal
      Suppose $X \sim (\mu, \sigma^2)$, then $aX + b \sim N(a\mu + b, a^2\sigma^2)$.
    - Reproductive property of Normal: Sum of normal distributions is still normal.

Central Limit Theorem

- Suppose $X_1 \ldots X_n$ are identical, independent distributed random variables with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$, so we have the sample mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{with } E[\bar{X}] = \mu \text{ and } Var(\bar{X}) = \frac{\sigma^2}{n}$$

  Thus, as $n \to \infty$, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$
- Same as let $X = \sum_{i=1}^{n} X_i$ with $E[X] = n\mu$ and $Var(X) = n\sigma^2$, in this case:

$$Y' = \frac{\bar{X} - \mu}{\sigma\sqrt{n}}$$

- Continuity Correction: when $X_1 \ldots X_n$ being estimated is discrete
  $P(x \geq 87) = P(x > 86.5)$ $\quad$ $P(x > 87) = P(x > 87.5)$
  $P(x \leq 87) = P(x < 87.5)$ $\quad$ $P(x < 87) = P(x < 86.5)$

Tail Bounds

- Markov's Inequality: $P(X \geq \alpha) \leq \frac{E[X]}{\alpha}$
- Chebyshev's Inequality: Suppose $E[Y] = \mu$ and $Var(Y) = \sigma^2$, then

$$P(|Y - \mu| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2} \text{ for any } \alpha \in R$$

- Chernoff's Bound: Suppose $X \sim Bin(n, p)$. Then for any $0 < \delta < 1$,
    - $P(X > (1 + \delta)\mu) \leq e^{-\frac{\delta^2\mu}{3}}$
    - $P(X > (1 - \delta)\mu) \leq e^{-\frac{\delta^2\mu}{2}}$

Law of Large Numbers

- Let $X_1 \ldots X_n$ be identical, independent distributed variables with common mean μ and variance σ^2. Let $\bar{X}$ be the sample mean for a sample size n. Then:
    - Weak Law of Large Numbers:

$$\text{for any } \epsilon > 0,\ \lim_{n\to\infty} P(|\overline{X_n} - \mu| > \epsilon) = 0$$

    - Strong Law of Large Numbers:

$$P(\lim_{n=\infty} \overline{X_n} = \mu) = 1$$

- The strong law implies the weak law but not vice versa.

Likelihood

- Realization/sample of a random variable: the actual values observed.
- Let $x_1 \dots x_n$ be realizations of random variable X, we define the likelihood function to be the probability of seeing these data:
  - If X is discrete with mass function $p_X(x|\theta)$:
  $$L(x_1 \dots x_n|\theta) = \Pi_{i=1}^{n} p_X(x_i|\theta)$$
  $$\ln L(x_1 \dots x_n|\theta) = \sum_{i}^{n} \ln p_X(x_i|\theta)$$
  - If X is continuous with density $f_X(x|\theta)$:
  $$L(x_1 \dots x_n) = \Pi_{i=1}^{n} f_X(x_i|\theta)$$
  $$\ln L(x_1 \dots x_n|\theta) = \sum_{i}^{n} \ln f_X(x_i|\theta)$$
- Maximum Likelihood Estimator (MLE): maximizes the likelihood function, denote as $\hat{\theta}$.
  Steps of finding MLE:
  - Find likelihood and log-likelihood of data
  - Take derivative of log-likelihood and find critical points
  - Use second derivative test to show $\hat{\theta}$ is a maximizer, that $\frac{\delta^2 L}{\delta \theta^2} < 0$ at $\hat{\theta}$, also check points of non-differentiability and boundary points.
- Bias: the bias of an estimator $\hat{\theta}$ for the true parameter θ is defined as
  $$Bias(\hat{\theta}, \theta) = E[\hat{\theta}] - \theta.$$
  An estimator is unbiased if and only if the bias of the estimator is 0.

Confidence Intervals

- $(\hat{\theta} - \Delta, \hat{\theta} + \Delta)$ is a $100(1-\alpha)\%$ confidence interval for θ if and only if
  $$P\left(\theta \in \left(\hat{\theta} - \Delta, \hat{\theta} + \Delta\right)\right) \geq 1 - \alpha.$$