1002 Midterm Cheat Sheet / Notes

USE AT YOUR OWN DISCRETION, THERE MIGHT BE TYPOS

October 2021

1 Probability

Conditional Probability:

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

Chain rule (2 events):

$$\begin{aligned} P(A \cap B) &= P(A)P(B|A) \\ &= P(B)P(A|B) \end{aligned} \tag{1}$$

DeMorgan's Law

More DeMorgan's Law

$$\begin{aligned} \text{One identity:} (A \cup B)^c &= A^c \cap B^c \\ \text{Another Identity} (A \cap B)^c &= A^c \cup B^c \end{aligned} \tag{2}$$

For two events

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

For three events

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

Chain rule (many events):

$$\begin{aligned} P(A \cap B \cap C) &= P(A)P(B \cap C|A) \\ &= P(A)P(B|A)P(C|A \cap B) \end{aligned} \tag{3}$$

Note the order of terms is arbitrary, can pull out B, or C first if we wanted. Consider what we know and choose wisely.

Law of Total Probability:

$$\begin{aligned} P(B) &= \sum_i P(B \cap A_i) \\ &= \sum_i P(A_i)P(B|A_i) \end{aligned} \tag{4}$$

Where $A_1, \ldots, A_n$ are disjoint events.

Bayes Rule:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{5}$$

Independence of Two Events

$$P(A|B) = P(A) \text{ and } P(A \cap B) = P(A)P(B) \tag{6}$$

Only if the two events A,B are independent.

Conditional Independence

$$P(A|B \cap C) = P(A|C) \text{ and } P(A \cap B|C) = P(A|C)P(B|C) \tag{7}$$

2 Modeling Discrete Data

Popular distributions

- Binomial(n,p) = $\binom{n}{k} \times p^k \times (1-p)^{n-k}$

  Good for knowing the probability of K Coin flips being Heads out of N. N is the number of times you are flipping the coin, p is the chance of success (heads).

- Geometric(a) = $\alpha \times (1-\alpha)^{n-1}$

  Assumes that all events are independent, all probabilities are the same. With this formula, you are asking: whats the probability of getting a streak of heads that ends on the a^{th} attempt? ($a-1$ heads, then 1 tails)

- Poisson Distribution:

$$P(\lambda) = \frac{\lambda^a e^{-\lambda}}{a!}$$

  that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event.

  a is the number of occurrences. IF you wanted to calculate $P(X >= 3) = 1 - P(0) - P(1) - P(2)$

  As n approaches infinity, a binomial random variable with parameters, n and $\frac{\lambda}{n}$ tends to the distribution of a poisson distribution with parameter λ

  For any period of time of length t, if t is small enough, the probability of an earthquake occurring in that period is equal to λ and the probability of more than one earthquake is negligible. λ is a fixed parameter representing the rate at which earthquakes occur.

  Each earthquake occurs independently from other earthquakes (we are ignoring aftershocks).

- **Negative Binomial**

$$\textbf{Negative Binomial } = \binom{k+r-1}{r} \times (1-p)^r \times p^k$$

  Intuition think of the dart problem on the homework.

Combinatorics formula

$$\frac{n!}{k!(n-k)!}$$

3 MLE For Discrete Distributions

Example 4.4 (Maximum-likelihood estimator for the Bernoulli distribution). Let $X := \{x_1, \ldots, x_n\}$ be n data points equal to zero or one, representing the occurrence of some event of interest. Assuming that the data are i.i.d., we decide to fit a Bernoulli model with parameter θ (in this case there is only one parameter). The likelihood function is equal to

$$\mathcal{L}_X(\theta) = \prod_{i=1}^{n} p_\theta(x_i) \tag{68}$$

$$= \theta^{n_1}(1-\theta)^{n_0}, \tag{69}$$

where n_0 and n_1 are the number of observations equal to zero and one respectively. The log-likelihood function equals

$$\log \mathcal{L}_X(\theta) = n_1 \log \theta + n_0 \log(1-\theta). \tag{70}$$

Figure 11 shows the likelihood and log-likelihood functions for $n_0 = 40$ and $n_1 = 60$.

The maximum-likelihood estimator of the parameter θ is

$$\theta_{\mathrm{ML}} = \arg\max_\theta \log \mathcal{L}_X(\theta) \tag{71}$$

$$= \arg\max_\theta n_1 \log \theta + n_0 \log(1-\theta). \tag{72}$$

The derivative and second derivative of the log-likelihood function equal

$$\frac{\mathrm{d} \log \mathcal{L}_X(\theta)}{\mathrm{d}\theta} = \frac{n_1}{\theta} - \frac{n_0}{1-\theta}, \tag{73}$$

$$\frac{\mathrm{d}^2 \log \mathcal{L}_X(\theta)}{\mathrm{d}\theta^2} = -\frac{n_1}{\theta^2} - \frac{n_0}{(1-\theta)^2} < 0 \quad \text{for all } \theta \in [0,1]. \tag{74}$$

The function is concave, as the second derivative is negative. This is good news, because it means that there cannot be different local maxima. The maximum is at the point where the first derivative equals zero, namely

$$\theta_{\mathrm{ML}} = \frac{n_1}{n_0 + n_1}. \tag{75}$$

The estimator is the fraction of samples that equal one. This is equivalent to estimating θ using the empirical probability of observing a one. △

Figure 1: Bernoulli MLE

MLE for Gaussian Distribution Use mean and std dev of empirical data
Uniform Distribution MLE a = min(data points), b = max(data points)

$$\log \mathcal{L}_{\{x_1,\ldots,x_n\}}(\theta) = \sum_{i=1}^{n} \log p_\theta(x_i) \tag{76}$$

$$= \sum_{i=1}^{n} \log\left(\theta^{x_i}(1-\theta)\right) \tag{77}$$

$$= \sum_{i=1}^{n} \left(x_i \log\theta + \log(1-\theta)\right) \tag{78}$$

$$= \left(\sum_{i=1}^{n} x_i\right) \log\theta + n\log(1-\theta) \tag{79}$$

$$= n_{\text{made}} \log\theta + n_{\text{missed}} \log(1-\theta), \tag{80}$$

where $n_{\text{made}} = \sum_{i=1}^{n} x_i$ is the number of made free throws and $n_{\text{missed}} = n$ is the number of missed free throws. The log-likelihood is exactly the same as the one from the Bernoulli model in Example 4.4. This makes sense, if we focus on individual free throws instead of on streaks of made free throws, our assumptions imply that the data are realizations of i.i.d. Bernoulli random variables with parameter θ. By the same argument as in Example 4.4, the maximum-likelihood estimator for θ equals the fraction of made free throws,

$$\theta_{\text{ML}} = \frac{n_{\text{made}}}{n_{\text{missed}} + n_{\text{made}}} \tag{81}$$

$$= 0.875. \tag{82}$$

Figure 2: Geomtric MLE

$$\mathcal{L}_X(\lambda) = \prod_{i=1}^{n} p_{\tilde{x}}(x_i) \tag{84}$$

$$= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}, \tag{85}$$

so the log-likelihood equals

$$\log \mathcal{L}_X(\lambda) = \sum_{i=1}^{n} \left(x_i \log\lambda - \lambda - \log(x_i!)\right). \tag{86}$$

The data is available here: http://iew3.technion.ac.il/serveng/callcenterdata.

The derivative and second derivative of the log-likelihood are

$$\frac{\mathrm{d}\log\mathcal{L}_X(\lambda)}{\mathrm{d}\lambda} = \sum_{i=1}^{n} \frac{x_i}{\lambda} - 1, \tag{87}$$

$$\frac{\mathrm{d}^2\log\mathcal{L}_X(\lambda)}{\mathrm{d}\lambda^2} = -\sum_{i=1}^{n} \frac{x_i}{\lambda^2} < 0. \tag{88}$$

The function is concave, as the second derivative is negative. The maximum is consequently at the point where the first derivative equals zero, namely

$$\lambda_{\text{ML}} = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{89}$$

Figure 3: Poisson Discrete MLE

4 Modeling Continuous Data

Cumulative Distribution Funciton

$$F_{\tilde{a}}(a) = P(\tilde{a} \leq a)$$

The probability that $\tilde{a} \leq a$. CDFs are always non-decreasing, and take values from [0,1]. Calculating a a probability of being in the the range of two values with cdfs:

$$\begin{aligned} P(a < \tilde{a} < b) &= F_{\tilde{a}}(b) - F_{\tilde{a}}(a) \\ &= \int_a^b f_{\tilde{a}}(a)da \end{aligned} \tag{8}$$

Probability Density Function

Captures the instantaneous rate of change of the random variable at that point, we obtain a PDF ($f_{\tilde{a}}$) by differentiating the CDF ($F_{\tilde{a}}$). Values in the PDF are not probabilities, but densities. You can have a value such that $f_{\tilde{a}} > 1$. For a pdf to be valid, it must have only non negative numbers and integrate to 1.

$$f_{\tilde{a}}(a) = \frac{dF_{\tilde{a}}(a)}{da}$$

Kernel Density Estimate

Definition 5.3 (Kernel density estimate). *Let $X := \{x_1, x_2, \ldots, x_n\}$ denote a real-valued dataset. The corresponding kernel density estimate with bandwidth h is*

$$f_{X,h}(a) := \frac{1}{n\,h} \sum_{i=1}^{n} K\left(\frac{a - x_i}{h}\right), \tag{80}$$

where K is a kernel function centered at the origin that satisfies

$$K(a) \geq 0 \quad \text{for all } a \in \mathbb{R}, \tag{81}$$

$$\int_{\mathbb{R}} K(a)\, dx = 1. \tag{82}$$

A popular choice for the kernel is the Gaussian function

$$K(a) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{a^2}{2}\right), \tag{83}$$

Figure 4: KDE

Detailed solution for 6a if we use Gaussian KDE

We went over how to do Gaussian KDE for estimating pdf in practice question 6a briefly in the lab. Here's the detailed solution:

To estimate the pdf we assume three Gaussian distributions $p_1 \sim \mathcal{N}(7.5, \sigma_1^2), p_2 \sim \mathcal{N}(10, \sigma_2^2), p_3 \sim \mathcal{N}(32.5, \sigma_3^2)$ for some standard deviations $\sigma_1, \sigma_2, \sigma_3$.

$p_1(t) = \frac{1}{\sqrt{2\pi}\sigma_1}\exp\big((t-7.5)^2/2\sigma_1^2\big)/C$

$p_2(t) = \frac{1}{\sqrt{2\pi}\sigma_2}\exp\big((t-10)^2/2\sigma_2^2\big)/C$

$p_3(t) = \frac{1}{\sqrt{2\pi}\sigma_3}\exp\big((t-32.5)^2/2\sigma_3^2\big)/C$

where C is the normalization constant. $C = \sum_{i=1}^{3}\int_{-\infty}^{\infty} p_i(t)dt = 3$

The pdf can be writen as $f_{\tilde{t}}(t) = \frac{1}{3}\Big[\frac{1}{\sqrt{2\pi}\sigma_1}\exp\big((t-7.5)^2/2\sigma_1^2\big) + \frac{1}{\sqrt{2\pi}\sigma_2}\exp\big((t-10)^2/2\sigma_2^2\big) + \frac{1}{\sqrt{2\pi}\sigma_3}\exp\big((t-32.5)^2/2\sigma_3^2\big)\Big]$

Figure 5: KDE Example from Lab

5 Continuous Distributions

Continuous Uniform Distribution

$$PDF = \frac{1}{b-a} \quad \text{where a is the min range, b the max range of function}$$

$$CDF = \begin{cases} 0 & for\ x < a \\ \frac{x-a}{b-a} & for\ x \in [a,b] \\ 1 & for\ x > b \end{cases}$$

Think of the uniform CDF as a straight line that connects 0 to 1 from a to b

Exponential Distribution PDF

An exponential random variable $\tilde{t}$ with parameter λ has a PDF:

$$PDF = f_{\tilde{t}}(t) = \begin{cases} \lambda e^{-\lambda t}, & if\ t \geq 0, \\ 0, & otherwise. \end{cases}$$

The exponential distribution is memoryless, such that when conditioning on a certain time $\tilde{t}|\tilde{t} > t_0$ we simply move over the exponential distribution $t - t_0$:

$$f_{\tilde{t}|\tilde{t}>t_0}(t) = \lambda e^{-\lambda(t-t_0)}$$

Exponential Distribution CDF

$$CDF = F_{\tilde{t}}(t) = \begin{cases} 1 - e^{-\lambda t}, & if\ t \geq 0, \\ 0, & otherwise. \end{cases}$$

Note if there's a scalar it gets preserved during integration. For example:

$$PDF = \alpha\lambda e^{-\lambda t} \rightarrow CDF = F_{\tilde{t}}(t) = \begin{cases} 1 - \alpha e^{-\lambda t}, & if\ t \geq 0, \\ 0, & otherwise. \end{cases}$$

Conditional CDF

$$Conditional\ CDF = F_{\tilde{t}|\tilde{t}>t_0}(t) = 1 - e^{-\lambda(t-t_0)}$$

Gaussian Distribution

Gaussian PDF with mean μ and variance σ^2 (N(mu,σ^2)):

$$PDF = f_{\tilde{a}}(a) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(a-\mu)^2}{2\sigma^2}}$$

MLE for Exponential

Lemma 7.3 (Maximum-likelihood estimator for the exponential distribution)**.** *Let $X := \{x_1, x_2, \ldots, x_n\}$ denote a dataset of nonnegative real-valued numbers. The maximum-likelihood estimator of the parameter of the exponential distribution under i.i.d assumptions equals*

$$\lambda_{\mathrm{ML}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i}. \tag{115}$$

Proof. The log-likelihood is given by

$$\log \mathcal{L}_X(\lambda) = \sum_{i=1}^{n} \log f_\lambda(x_i) \tag{116}$$

$$= \sum_{i=1}^{n} \log \lambda \exp(-\lambda x_i) \tag{117}$$

$$= n \log \lambda - \lambda \sum_{i=1}^{n} x_i. \tag{118}$$

The derivative and second derivative of the log-likelihood function are given by

$$\frac{\mathrm{d} \log \mathcal{L}_{x_1,\ldots,x_n}(\lambda)}{\mathrm{d}\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i, \tag{119}$$

$$\frac{\mathrm{d}^2 \log \mathcal{L}_{x_1,\ldots,x_n}(\lambda)}{\mathrm{d}\lambda^2} = -\frac{n}{\lambda^2} < 0 \quad \text{for all } \lambda > 0. \tag{120}$$

The function is concave, as the second derivative is negative, so there cannot be different local maxima. The maximum is obtained by setting the first derivative to zero. □

Inverse Transform Sampling

Two reasons:

1. Generating uniform samples from the unit interval [0,1]
2. Transforming the uniform samples so they have the desired distribution

Algorithm 8.1 (Inverse-transform sampling). *Let $\tilde{a}$ be a continuous random variable with cdf $F_{\tilde{a}}$ and $\tilde{u}$ a random variable that is uniformly distributed in $[0, 1]$ and independent of $\tilde{a}$.*

1. *Obtain a sample u of $\tilde{u}$.*
2. *Set $x := F_{\tilde{a}}^{-1}(u)$.*

Example 8.3 (Sampling from an exponential distribution). Let $\tilde{a}$ be an exponential random variable with parameter λ. Its cdf $F_{\tilde{a}}(x) := 1 - e^{-\lambda x}$ is invertible in $[0, \infty]$. Its inverse equals

$$F_{\tilde{a}}^{-1}(u) = \frac{1}{\lambda} \log\left(\frac{1}{1-u}\right). \tag{138}$$

Figure 6: Inverse Transform Sampling

Inverse Transform Sampling Algorithm: when calculating the inverse cdf, integrate from the bottom range of each pdf to a random variable $\tilde{t}$. Make sure, that if you have multiple pdfs (and are calculating multiple cdfs), you add the cumulative value after integrating. For example, if i have two pdfs, and the first cdf covers one half of the probability, then after integrating the second pmf to find the second cdf, add the value of 1/2 to the result. Then you can inverse by switching y,t to calculate inverse. Then plug in your sample according to the sample value (if I have a sample value $< 1/2$ I would use the first cdf, if its greater I would use the second one).

Intuition: You have two use cases for inverse transform sampling:

- You plug in a sample from a uniform distribution into a inverse cdf to get the corresponding pdf/cdf value that corresponds to the percentile of the sample from the uniform. I.e. You want to find the score you would need to be in the 90th percentile. You plug in .9 to the inverse cdf and you get the score you would need.
- The inverse is true. You can plug a pmf value into the inverse cdf to find the according percentile (uniform distribution sample).

6 Modeling Multivariate Discrete Data

Joint PDFs Notation:

$$P_{\tilde{a},\tilde{b}}(a,b) = P(\tilde{a} = a, \tilde{b} = b) \qquad a \in R_{\tilde{a}} \;\; b \in R_{\tilde{b}}$$

Joint pdfs must be non-negative (they represent probabilities), and must sum to 1.

$$\sum_{a \in R_{\tilde{a}}} \sum_{b \in R_{\tilde{b}}} p_{\tilde{a},\tilde{b}}(a,b) = 1$$

Marginal PMF

$$p_{\tilde{a}}(a) = \sum_{b \in B} p_{\tilde{a},\tilde{b}}(a,b)$$

Naive Bayes

Here we are trying to predict a classifier y given some data that has features $x_1, x_2, \ldots$. To keep things simple, in the formula below were assuming y is a binary class (i.e. 1 is republican 0 is democrat) and we only have 3 features x_1, x_2, x_3

$$Naive\ Bayes = P(y|x_1 \cap x_2 \cap x_3) = \frac{P(y \cap x_1 \cap x_2 \cap x_3)}{P(x_1 \cap x_2 \cap x_3)} = \frac{P(y)P(x_1|y)P(x_2|y)P(x_3|y)}{\sum_{y=0}^{1} P(y \cap x_1 \cap x_2 \cap x_3)}$$

$$= \frac{P(y=1)P(x_1|y=1)P(x_2|y=1)P(x_3|y=1)}{P(y=1)P(x_1|y=1)P(x_2|y=1)P(x_3|y=1) + P(y=0)P(x_1|y=0)P(x_2|y=0)P(x_3|y=0)}$$

The features x_1, x_2, x_3 should all be set to some value as well, I just didn't have the space to do such notation.

Miscellaneous, (Geometric Series, Derivatives and Integrals)

Geometric Series

$$\sum_{i=1}^{k} r^i = \frac{r(1-r^i)}{1-r} \tag{9}$$

Note if r is a fraction, as k approaches infinity $r^i = 0$ thus the top half of the equation will be just $r(1-0) = r$

Derivative	Integral (Antiderivative)
$\frac{d}{dx} n = 0$	$\int 0 \, dx = C$
$\frac{d}{dx} x = 1$	$\int 1 \, dx = x + C$
$\frac{d}{dx} x^n = nx^{n-1}$	$\int x^n \, dx = \frac{x^{n+1}}{n+1} + C$
$\frac{d}{dx} e^x = e^x$	$\int e^x \, dx = e^x + C$
$\frac{d}{dx} \ln x = \frac{1}{x}$	$\int \frac{1}{x} \, dx = \ln x + C$
$\frac{d}{dx} n^x = n^x \ln x$	$\int n^x \, dx = \frac{n^x}{\ln n} + C$
$\frac{d}{dx} \sin x = \cos x$	$\int \cos x \, dx = \sin x + C$
$\frac{d}{dx} \cos x = -\sin x$	$\int \sin x \, dx = -\cos x + C$
$\frac{d}{dx} \tan x = \sec^2 x$	$\int \sec^2 x \, dx = \tan x + C$
$\frac{d}{dx} \cot x = -\csc^2 x$	$\int \csc^2 x \, dx = -\cot x + C$
$\frac{d}{dx} \sec x = \sec x \tan x$	$\int \tan x \sec x \, dx = \sec x + C$
$\frac{d}{dx} \csc x = -\csc x \cot x$	$\int \cot x \csc x \, dx = -\csc x + C$
$\frac{d}{dx} \arcsin x = \frac{1}{\sqrt{1-x^2}}$	$\int \frac{1}{\sqrt{1-x^2}} \, dx = \arcsin x + C$
$\frac{d}{dx} \arccos x = -\frac{1}{\sqrt{1-x^2}}$	$\int -\frac{1}{\sqrt{1-x^2}} \, dx = \arccos x + C$
$\frac{d}{dx} \arctan x = \frac{1}{1+x^2}$	$\int \frac{1}{1+x^2} \, dx = \arctan x + C$
$\frac{d}{dx} \operatorname{arc\,cot} x = -\frac{1}{1+x^2}$	$\int -\frac{1}{1+x^2} \, dx = \operatorname{arc\,cot} x + C$
$\frac{d}{dx} \operatorname{arc\,sec} x = \frac{1}{x\sqrt{x^2-1}}$	$\int \frac{1}{x\sqrt{x^2-1}} \, dx = \operatorname{arc\,sec} x + C$
$\frac{d}{dx} \operatorname{arc\,csc} x = -\frac{1}{x\sqrt{x^2-1}}$	$\int -\frac{1}{x\sqrt{x^2-1}} \, dx = \operatorname{arc\,csc} x + C$

Figure 7: Common Derivative and Integral Formula

Rule name	Rule	Example
Product rule	$\ln(x \cdot y) = \ln(x) + \ln(y)$	$\ln(3 \cdot 7) = \ln(3) + \ln(7)$
Quotient rule	$\ln(x / y) = \ln(x) - \ln(y)$	$\ln(3 / 7) = \ln(3) - \ln(7)$
Power rule	$\ln(x^y) = y \cdot \ln(x)$	$\ln(2^8) = 8 \cdot \ln(2)$
ln derivative	$f(x) = \ln(x) \Rightarrow f'(x) = 1 / x$	
ln integral	$\int \ln(x)dx = x \cdot (\ln(x) - 1) + C$	
ln of negative number	$\ln(x)$ is undefined when $x \leq 0$	
ln of zero	$\ln(0)$ is undefined	
	$\lim_{x \to 0^+} \ln(x) = -\infty$	
ln of one	$\ln(1) = 0$	
ln of infinity	$\lim \ln(x) = \infty$,when $x \to \infty$	
Euler's identity	$\ln(-1) = i\pi$	

Figure 8: Properties of Natural Logarithm