Midterm exam, 機器學習, Fall 2020. Open book but no calculators/cell phones allowed. Answers may include e^2, $\sqrt{2}$, etc. but simplify when possible.

Your Name: ____________________

Problem 1.
Suppose
1. we have some observed data X (a set of real numbers $\{x_1, ..., x_n\}$
2. We assume the data are random samples generated by a normal distribution of unknown mean μ.

Question 1a
What is the maximum likelihood estimator for the μ?

Solution: The arithmetic mean, $\mu = \hat{x} \stackrel{\text{def}}{=} \Sigma_i x_i / n$.

Question 1b
Given $\mu = 0$, what is the maximum likelihood estimator for σ^2?

Give the mathematical derivation for your answers:

Solution:

MLE of Normal Mean

We need to maximize the likelihood with respect to μ. The easiest way is to take the derivative of the log likelihood with respect to μ and solve for its zero value. For variety, I show a different way.

$$\arg\max_{\mu} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{\sigma^2}\right))$$

$$= \arg\max_{\mu} \prod_{i=1}^{n} \exp\left(-\frac{(x_i-\mu)^2}{\sigma^2}\right)$$

$$= \arg\max_{\mu} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^2}\right)$$

$$= \arg\max_{\mu} \sum_{i=1}^{n} -\frac{(x_i-\mu)^2}{\sigma^2} \qquad \text{exp function is monotonically increasing}$$

$$= \arg\min_{\mu} \sum_{i=1}^{n} (x_i-\mu)^2 \;=\; \arg\min_{\mu} \left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\mu + \sum_{i=1}^{n} \mu^2)\right)$$

$$= \arg\min_{\mu} \left(\sum_{i=1}^{n} \mu^2 - \sum_{i=1}^{n} 2\mu x_i\right) \;=\; \arg\min_{\mu} \left(n\mu^2 - (\sum_{i=1}^{n} x_i)2\mu\right)$$

$$= \arg\min_{\mu} (n\mu^2 - n\hat{x}2\mu) \;=\; \arg\min_{\mu} (\mu^2 - \hat{x}2\mu + \hat{x}^2) \qquad \text{added constant } \hat{x}^2$$

$$= \arg\min_{\mu} (\mu - \hat{x})^2 = \hat{x} \checkmark$$

Solution:

MLE of Normal Variance

Method 1: treating the log likelihood as a function of σ

The likelihood is:

$$\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp(\frac{-(x_i-0)^2}{2\sigma^2})$$

It is convenient to work with the log likelihood, which has the same maximum as the likelihood.

$$\begin{aligned}
&\max_\sigma \left(\sum_{i=1}^{n} \left(\frac{-(x_i-0)^2}{2\sigma^2} - \lg(\sigma) - \lg(\sqrt{2\pi}) \right) \right) \\
&= \max_\sigma \left(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (x_i-0)^2 \quad - n\lg(\sigma) - n\lg(\sqrt{2\pi}) \right) \\
&= \max_\sigma \left(\frac{-\sum_i (x_i^2)}{2\sigma^2} - n\lg(\sigma) \right) && \text{log likelihood}(\ \sigma\ )
\end{aligned}$$

Let S_2 denote $\sum_i x_i^2$. Take derivative and solve for the value of σ which makes it equal to zero.

$$\begin{aligned}
\text{find } \sigma:\ & \frac{S_2}{\sigma^3} - \frac{n}{\sigma} = 0 && \text{derivative } \frac{d}{d\sigma} \text{ of log likelihood}(\sigma) \\
& \frac{S_2}{\sigma^2} - n = 0 \Rightarrow \quad \frac{S_2}{\sigma^2} = n \Rightarrow \quad \sigma^2 = \frac{S_2}{n} \checkmark
\end{aligned}$$

Method 2: treating the log likelihood as a function of σ^2

In this method, I find it more clear to use a plain (not squared) variable to denote σ^2, so let $y := \sigma^2$.

$$\max_y \left(\sum_{i=1}^{n} \frac{-(x_i-0)^2}{2y} - \lg(\sqrt{y}) - \lg(\sqrt{2\pi}) \right) = \max_y \left(\frac{-\sum_i (x_i^2)}{2ny} - \lg(\sqrt{y}) \right) \qquad \text{log likelihood}(y)$$

Again, let S_2 denote $\sum_i x_i^2$. Take derivative and solve for the value of y which makes it equal to zero.

$$\begin{aligned}
\text{find } y:\ & \frac{S_2 n}{2y^2} - \frac{1}{2y} = 0 && \text{derivative } \frac{d}{dy} \text{ of log likelihood}(y) \\
& \frac{S_2 n}{2y} - \tfrac{1}{2} = 0y \Rightarrow \quad S_2 n - y = 0 \Rightarrow \quad y = \frac{S_2}{n} \Rightarrow \sigma^2 = \frac{S_2}{n} \checkmark
\end{aligned}$$

Your Name: ______________________________

Problem 2.
Let v be a random variable defined by these values and probabilities.

v	probability
2	0.4
3	0.3
4	0.2
5	0.1

Let $V(n) = v_1 + v_2 + \cdots + v_n$ be the sum of n independent samples of v.

Question 2
Derive the mean and standard deviation of $V(n)$.

Solution: First we manually compute the mean and variance of a single sample v. Then we use "the expectation of a sum of $f(x_i)$ is the sum of the expectation of $f(x_i)$" property to extend that result to $V(n)$. Let μ_i and σ_i^2 denote the mean and variance of a single sample v_i.

$$\text{mean } (v_i) =: \mu_i \stackrel{\text{def}}{=} \mathrm{E}[v] = 0.4 \cdot 2 + 0.3 \cdot 3 + 0.2 \cdot 4 + 0.1 \cdot 5 = 3$$

$$\text{variance } (v_i) =: \sigma_i^2 \stackrel{\text{def}}{=} \mathrm{E}[(v_i - \mu_i)^2] = 0.4 \cdot (2\text{-}3)^2 + 0.3 \cdot (3\text{-}3)^2 + 0.2 \cdot (4\text{-}3)^2 + 0.1 \cdot (5\text{-}3)^2 = 1$$

$$\text{mean } V(n) =: \mu_n \stackrel{\text{def}}{=} \mathrm{E}\left[\sum_{i=1}^{n} v_i\right] = \sum_{i=1}^{n} \mathrm{E}[v] = \sum_{i=1}^{n} \mu_i = n\mu_i = 3n$$

$$\text{variance } V(n) =: \sigma_n^2 \stackrel{\text{def}}{=} \mathrm{E}\left[\sum_{i=1}^{n} (v_i - \mu_i)^2\right] = \sum_{i=1}^{n} \left[\mathrm{E}[(v_i - \mu_i)^2]\right] = n\sigma_i^2 = n$$

$$\text{Std Dev } V(n) \stackrel{\text{def}}{=} \sqrt{\sigma_n^2} = \sqrt{n}$$

Your Name: ______________________________

Problem 3.
The Poisson distribution has parameter $\lambda \geqq 0$, defining a probability distribution over the non-negative integers $(0, 1, ...)$ as follows:

$$\text{Pois}(k; \lambda) \quad \stackrel{\text{def}}{\equiv} \quad P[k] = \frac{\lambda^k}{k! \exp(\lambda)}, \quad k \in \mathbb{N}_0$$

This problem involves inference from data generated by one of two Poisson distributions: $\text{Pois}(\lambda_1)$ or $\text{Pois}(\lambda_2)$. The following experiment is done.
1. λ is set to $\{\lambda_1, \lambda_2\}$ with probability m_1 and $m_2 = 1 - m_1$.
2. A random sample y is drawn from $\text{Pois}(\lambda)$

Question 3a
What is the posterior probability $\text{P}[\lambda = \lambda_1 | y = k]$?

Solution:

$$\begin{aligned}
\text{P}[\lambda = \lambda_1 | y = k] &= \frac{\text{P}[\lambda = \lambda_1]\text{P}[y|\lambda = \lambda_1]}{\text{P}[y = k]} \\
&= \frac{m_1 \frac{\lambda_1^y}{y! \exp(\lambda_1)}}{m_1 \frac{\lambda_1^y}{y! \exp(\lambda_1)} + m_2 \frac{\lambda_2^y}{y! \exp(\lambda_2)}} \\
&= \frac{m_1 \frac{\lambda_1^y}{\exp(\lambda_1)}}{m_1 \frac{\lambda_1^y}{\exp(\lambda_1)} + m_2 \frac{\lambda_2^y}{\exp(\lambda_2)}} \\
&= \frac{1}{1 + \frac{m_2}{m_1} \left(\frac{\lambda_2}{\lambda_1}\right)^y \exp(\lambda_1 - \lambda_2)}
\end{aligned}$$

Or equivalently,

$$\text{Posterior odds } \lambda_1 : \lambda_2 = m_1 \lambda_2^y \exp(\lambda_1) : m_2 \lambda_1^y \exp(\lambda_2)$$

Question 3b:
What kind of prior is this? Is it conjugate? Why or why not?

Solution: The prior on λ is a simple 2-value distribution $\lambda = \lambda_1$ or $\lambda = \lambda_2$ with odds $m_1 : m_2$. Could also be considered a very simple mixture model weighted by $m_1 : m_2$. The components being the trivial constant probability distributions: with probability 100%, $\lambda = \lambda_1$ or $\lambda = \lambda_2$ respectively.
The posterior is indeed conjugate as it has the same form (2-value distribution $\lambda = \lambda_1$ or $\lambda = \lambda_2$) but with weights: $m_1 \lambda_2^k \exp(\lambda_1) : m_2 \lambda_1^k \exp(\lambda_2)$
Note that the definition in this problem also defines a predictive distribution over k, and this distribution is a mixture model of two poisson distributions. After observing $k = y$, and updated posterior predictive is still a mixture model of the same two poisson distributions (but with different weights).

Your Name: ____________________________________

Problem 4.
A standard poker deck has 52 cards. 13 each of: ♠ ♣ ♥ ♦ . The entropy of a single card drawn at random is ≈ 5.7 bits of information. You cannot see the card, but I can.

Question 4a
If I told you the card is black (i.e. '♠' or '♣'); how much entropy would remain? (give numerical answer and reason)

Solution: The card is equally likely to be black or not, so the answer to that question gives us one bit of information. Therefore the remaining entropy $\lg(52) - 1 \approx 4.7$.

Question 4b
If I then told you the card was a spade '♠', how much entropy would remain then? (give numerical answer and reason)

Solution: Given the card is black, again the card is equally likely to be '♠' or not, so the answer to that question gives us one bit of information. Therefore the remaining entropy after the second answer is $\lg(52) - 1 - 1 \approx 3.7$.

Question 4c
Two cards are drawn from a fresh deck of cards. Let S_1, S_2 denote the first and second cards respectively. What is the mutual information $\mathrm{I}(S_1, S_2)$ (answer can include lg symbol).

Solution: $\mathrm{I}(S_1, S_2) = \mathrm{H}(S_2) - \mathrm{H}(S_2|S_1) = \lg(52) - \lg(51)$
Since $\mathrm{H}(S_2|S_1)$ is the entropy over 51 equally likely cards.

Your Name: ______________________________

Problem 5. Consider a classification problem with two features $F1 \in \{\texttt{0},\texttt{1},\texttt{2}\}$, $F2 \in \{\texttt{0},\texttt{1},\texttt{2},\texttt{3}\}$, Assume we know the two classes occur with equal probability: $\mathrm{P}[C = \texttt{A}] = \mathrm{P}[C = \texttt{B}] = 0.5$ (so you do not need to estimate $\mathrm{P}[C = \texttt{A}]$, just take it as given to be 0.5).

Training Data

F1	F2	Class
2	3	A
0	0	A
2	2	A
2	0	A
2	0	B
2	0	B
1	0	B
2	0	B

Test Data		P[C=A|F1,F2]:P[C=B|F1,F2] Using Prior:		
F1	F2	MLE	Jeffreys	Laplace
2	3			
1	1			
0	3			
1	0			
0	0			
1	3			
0	2			
0	1			

The above table gives the P[F|C] probabilities for each feature and class.

Question 5
Compute the probability a Naïve Bayes classifier would assign to $\mathrm{P}[C = \texttt{A}]$, using maximum estimation, Jeffrey's priors or Laplace priors respectively when estimating probabilities involving feature values. You may report the answer in terms of odds, so for example, if the $\mathrm{P}[C = \texttt{A}] = \frac{1}{3}$, you can report that as 1:2 (hint: it is easier to work with odds).
Given the Naïve Bayes assumption and that we know $\mathrm{P}[C = \texttt{A}] = \mathrm{P}[C = \texttt{B}]$, the most convenient way to compute the posterior odds uses the following relationships.

$$\mathrm{P}[C = \texttt{A}|F1, F2] : \mathrm{P}[C = \texttt{B}|F1, F2] \quad = \quad \mathrm{P}[F1|C = \texttt{A}] : \mathrm{P}[F2|C = \texttt{B}]$$

So first I would tally value counts for F1, and F2 in each class, then add in "pseudocounts" as appropriate for the given prior. Worksheet for intermediate calculations.

	Feature F1			Feature F2				Class
Value	**0**	**1**	**2**	**0**	**1**	**2**	**3**	
	1	0	3	2	0	1	1	A
counts	0	1	3	4	0	0	0	B
"counts"	1.5	0.5	3.5	2.5	0.5	1.5	1.5	A
Jeffreys	0.5	1.5	3.5	4.5	0.5	0.5	0.5	B
2×counts	3	1	7	5	1	3	3	A
Jeffreys	1	3	7	9	1	1	1	B
	2	1	4	3	1	2	2	A
Laplace	1	2	4	5	1	1	1	B
	1	3	4	3	1	1	3	A
TRUE	4	3	1	2	2	2	2	B

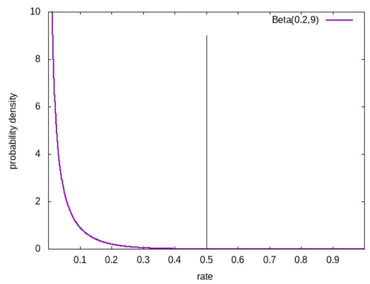
Solution:

Test Values		P[C=A|F1,F2]:P[C=B|F1,F2] Using Prior:		
F1	F2	MLE	Jeffreys	Laplace
2	3	$3:0=\infty$	$3:1$ (75%)	$2:1$ (67%)
1	1	$0:0=NaN$	$1:3$ (25%)	$1:2$ (33%)
0	3	$1:0=\infty$	$9:1$ (90%)	$4:1$ (80%)
1	0	$0:4=0$	$5:27$ (16%)	$3:10$ (23%)
0	0	$2:0=\infty$	$5:3$ (62%)	$6:5$ (55%)
1	3	$0:0=NaN$	$1:1$ (50%)	$1:1$ (50%)
0	2	$1:0=\infty$	$9:1$ (90%)	$4:1$ (80%)
0	1	$0:0=\infty$	$3:1$ (75%)	$2:1$ (67%)

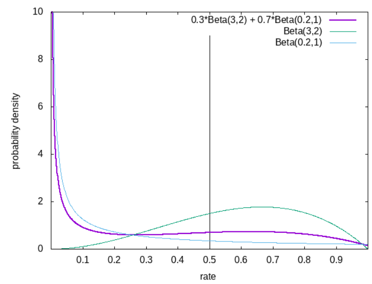
Your Name: ____

Problem 6.

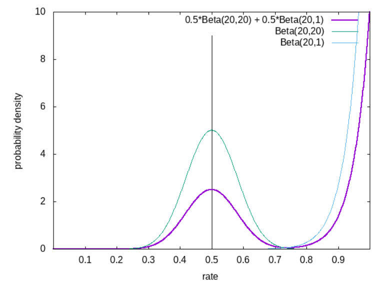
A



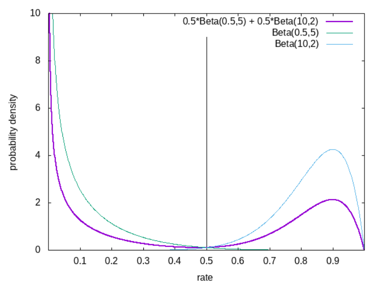
B



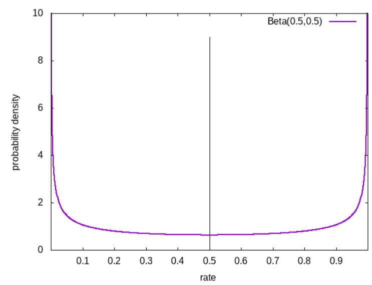
C



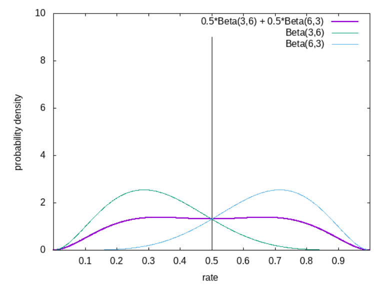
D



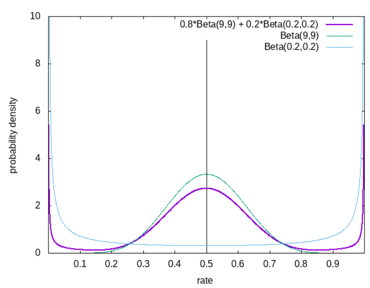
E



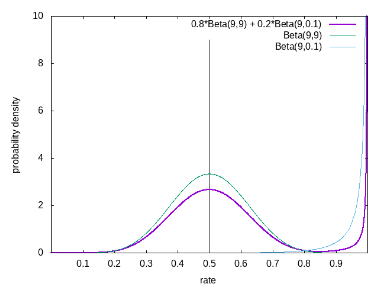
F



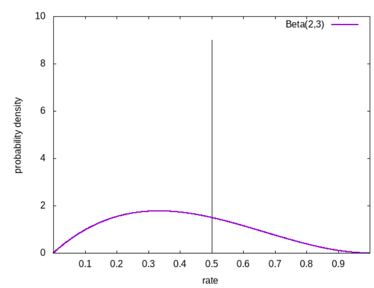
G



H



I



Question 6
What parameter values in the table on the previous path match which plot in the table above? Fill in the "Label" column and use the "Comment" column and/or space at bottom to explain your answers.

ID	Distribution	Label (A–I)	Comment
1	$0.8\,\text{Beta}(9,9) + 0.2\,\text{Beta}(9,0.1)$	H	main component near 0.5, another one at extreme rate=1
2	$\text{Beta}(0.5,0.5)$	E	Jeffrey Prior, symmetric with weight at extremes
3	$0.5\,\text{Beta}(3,6) + 0.5\,\text{Beta}(6,3)$	F	symmetric sum of two bell shapes, centered on $\frac{1}{3}$ and $\frac{2}{3}$
4	$\text{Beta}(2,3)$	I	bell shaped somewhat favoring rate < 0.5
5	$0.5\,\text{Beta}(20,20) + 0.5\,\text{Beta}(20,1)$	C	mixture of rate close to 0.5 or very near 1
6	$0.3\,\text{Beta}(3,2) + 0.7\,\text{Beta}(0.2,1)$	B	mixture of sharpish on near 0 and bell on ${}^{3}/_{5}$
7	$\text{Beta}(0.2,9)$	A	Extreme favoring of rate near zero
8	$0.5\,\text{Beta}(0.5,5) + 0.5\,\text{Beta}(10,2)$	D	Extreme towards 1 mixed with rounded on ${}^{5}/_{6}$
9	$0.8\,\text{Beta}(9,9) + 0.2\,\text{Beta}(0.2,0.2)$	G	Symmetric. Strong fair bell mixed with very sharp at extremes

Most of the plots above are unique enough to be easy to match to their parameters. The most similar ones are perhaps, C, G & H. But G is symmetric, while the other two are not. The component of H near rate=1 is much sharper than that of C, so that component should have a small value for α; and in fact it is 0.1 for H, compared to 1 for C.