Midterm exam for Fundamentals of Statistical Machine Learning (201910). Open book test.

Your Name: _____

1. Consider a computer program C which outputs '`a`' or '`b`' each time it runs. It is known that C uses a random generator to output `a` with probability p_a.

   No one knows the value of p_a, but you have a friend, Thomas, who plans to run the program some number of times and then predict the output of the next run of the program.

   Thomas loves Bayesian statistics, so his posterior probability estimate certainly will include a prior.

   Unfortunately Thomas is coy and will not tell you what his prior is.

   Instead he tells you that if he were to run the program 4 times and get an `a` each time, he would be 90% sure that next run would also output an `a`. Likewise, if he were to get 4 out of 4 `b`'s, he would be 90% sure that the next run would output another `b`.

   Question: Is Thomas using a uniform prior? Explain your answer.

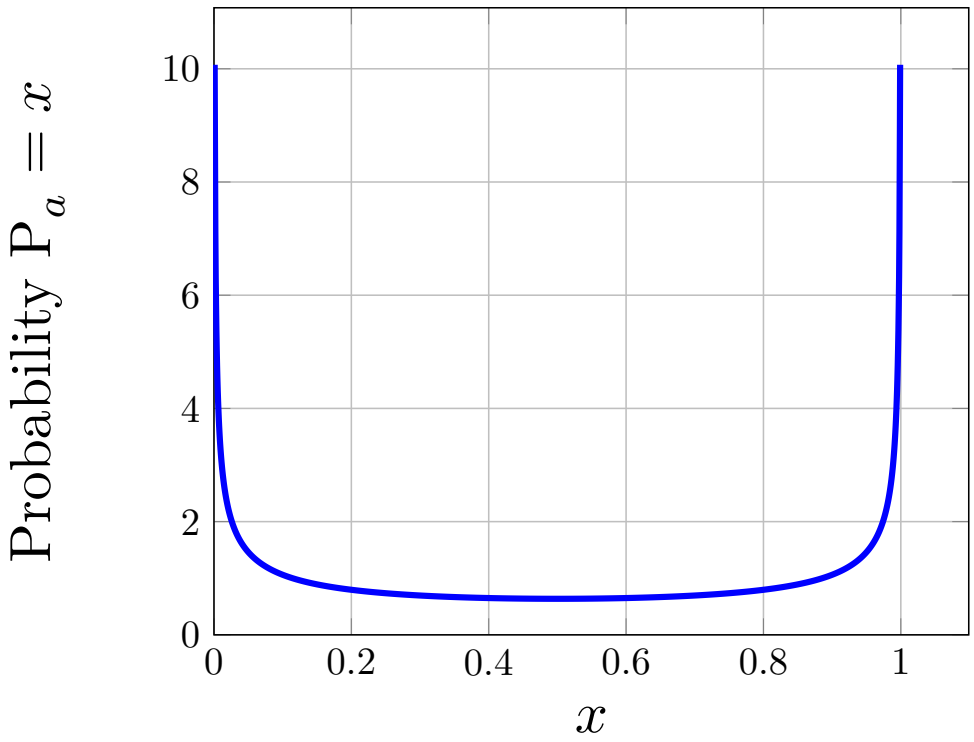


Figure 1: Beta(½,½) — Jeffrey's Prior

Solution:

Thomas is using a Jeffreys Prior, which is not the same as a uniform prior. A uniform prior stipulates a pseudocount of one for each class (in this case `a` or `b`), so after 4 `a`'s in a row, the posterior estimate of the probability of getting another `a` would be $\frac{4+1}{4+2} = 5/6 \neq 90\%$.

Instead Thomas is using a pseudocount of $1/2$ for both `a` and `b`, yielding $\frac{4^1/_2}{4+1} = 90\%$. This is the Jeffreys prior.

$$\text{Beta}(x|a,b) \stackrel{\text{def}}{=} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

A uniform prior is equivalent to $\text{Beta}(1,1) = 1x^{1-1}(1-x)^{1-1} = 1$.

The Jeffreys prior is $\text{Beta}(1/2, 1/2)$ with pdf:

$$\text{Beta}(1/2, 1/2) = \frac{\Gamma(1/2+1/2)}{\Gamma(1/2)\Gamma(1/2)} x^{1/2-1}(1-x)^{1/2-1} = \frac{1}{\sqrt{(\pi)}\sqrt{(\pi)}} x^{-\frac{1}{2}}(1-x)^{-\frac{1}{2}} = \frac{1}{\pi\sqrt{x(1-x)}}$$

Your Name: _____

2. This question asks about the entropy of a probability distribution in which two categories have been joined together.

   Let F denote a probability distribution over the first k natural numbers: $1, 2, \ldots, k$. Denote the probability $\mathrm{F}(i)$ as p_i, so $\sum_1^k p_i \equiv 1$.

   (i.e. F is like a die (骰子) with k-sides, each side having its own probability p_i)

   Let F′ be a probability distribution over $1, 2, \ldots, k-1$;
   almost the same as F, but with the last two elements ($k-1$ and k) merged.
   So, $p'_i = p_i$ $1 \leq i < k-1$, and $p'_{k-1} = p_{k-1} + p_k$. There is no p'_k (or equivalently $p'_k \stackrel{\text{def}}{=} 0$).

   Let H() denote information theoretic entropy.

   GIVEN: We know the theoretic entropy of F, $\mathrm{H}(\mathrm{F}) = 5.0$; and that $p_{k-1} = p_k = 0.01$.

   1. Give a lower bound on k and explain why k must be a least that large.
   2. What is $\mathrm{H}(\mathrm{F}')$, the entropy of F′?

Solution:

1. An entropy of h bits is the amount of information needed to select among 2^h equally likely things. So one way to have $\mathrm{H}(\mathrm{F}) = 5$ is the uniform distribution over $1, 2, \ldots, 32$. Therefore k could be as small as 32. Moreover for a fixed k, a uniform distribution maximizes entropy, so for example if k were 31, the entropy would have to be less than $lg\, 31 \approx 4.95 < 5$. Thus 32 is a lower bound on k.

2. Using the "increases incrementally when splitting categories" property of entropy,
We have:

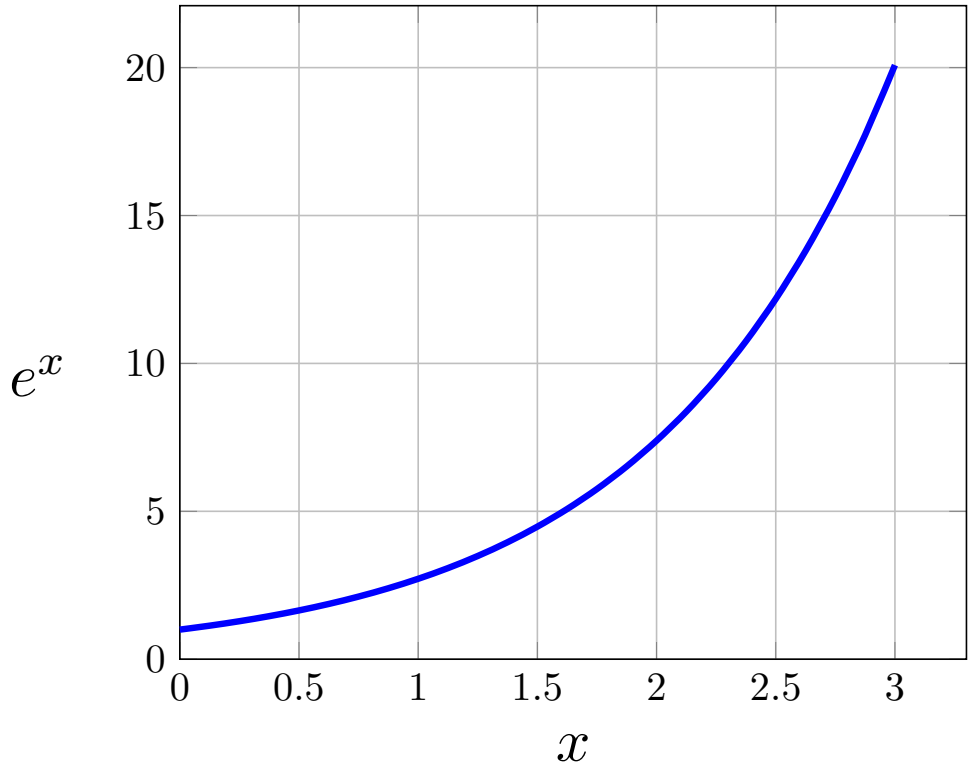
$$\mathrm{H}(\mathrm{F}) = \mathrm{H}(\mathrm{F}') + p'_{k-1}\mathrm{H}(0.5, 0.5) \tag{1}$$

Rearranging,

$$\mathrm{H}(\mathrm{F}') = \mathrm{H}(\mathrm{F}) - p'_{k-1}\mathrm{H}(0.5, 0.5) = 5 - 0.02 \cdot 1 = 4.98 \tag{2}$$

Because H(0.5,0.5) is a 50%–50% split, which is one bit of information.

Your Name: _____



3.

Consider a mixture model of two normal (one-dimensional Gaussian) distributions:

$$\mathrm{M}(x) = {}^{2}\!/_{3}\mathrm{N}_{\mathrm{a}} + {}^{1}\!/_{3}\mathrm{N}_{\mathrm{b}}$$

Denoted $\mathrm{N}_{\mathrm{a}} : \mathrm{N}(\mu_{\mathrm{a}}, \sigma_{\mathrm{a}}^2)$ and $\mathrm{N}_{\mathrm{b}} : \mathrm{N}(\mu_{\mathrm{b}}, \sigma_{\mathrm{b}}^2)$.

In other words, when generating points from M:
first one of the two components N_{a} or N_{b} are selected are random with N_{a} selected ${}^{2}\!/_{3}$ of the time,
and then a point is generated according to the mean and variance of the selected component.

A data point x_1 is sampled from $\mathrm{M}(x)$.

Consider the probability that x_1 "came from" N_{a}.

1. Give the general formula for that probability (a formula including x_1,...)
2. Give the approximate numerical value for the probability for the case:

x_1	μ_{a}	σ_{a}	μ_{b}	σ_{b}
4	-1	2	18	4

   Approximation should be made *without* using a caculator (不可以用計算機) and should be within ±10%.
   The answer can be given as a fraction, e.g. ${}^{2}\!/_{7}$ instead of 0.285714 ....

Solution:

For convenience, first define $z_{\rm a} = {}^{(\mu_{\rm a}-x_1)}/_{\sigma_{\rm a}}$, $z_{\rm b} = {}^{(\mu_{\rm b}-x_1)}/_{\sigma_{\rm b}}$;
and compute the probability ratio $\frac{\mathrm{P}[\mathrm{N_a}|x]}{\mathrm{P}[\mathrm{N_b}|x]}$ first.

General Formula

$$\frac{\mathrm{P}[\mathrm{N_a}|x]}{\mathrm{P}[\mathrm{N_b}|x]} = \frac{\mathrm{P}[\mathrm{N_a}]{}^1/_{\sigma_{\rm a}}\, EXP[-z_{\rm a}^2/2]}{\mathrm{P}[\mathrm{N_b}]{}^1/_{\sigma_{\rm b}}\, EXP[-z_{\rm b}^2/2]} = \frac{\mathrm{P}[\mathrm{N_a}]\sigma_{\rm b}\, EXP[z_{\rm b}^2/2]}{\mathrm{P}[\mathrm{N_b}]\sigma_{\rm a}\, EXP[z_{\rm a}^2/2]} = \frac{\mathrm{P}[\mathrm{N_a}]}{\mathrm{P}[\mathrm{N_b}]}\frac{\sigma_{\rm b}}{\sigma_{\rm a}}\, EXP[\frac{z_{\rm b}^2 - z_{\rm a}^2}{2}]$$

Finally convert the ratio into a probability via:

$$\mathrm{P}[\mathrm{A}] \equiv \frac{1}{1+\frac{\mathrm{P}[\mathrm{B}]}{\mathrm{P}[\mathrm{A}]}} \text{ obtaining } \mathrm{P}[\mathrm{N_a}|x] = \frac{1}{1+\frac{\mathrm{P}[\mathrm{N_b}]}{\mathrm{P}[\mathrm{N_a}]}\frac{\sigma_{\rm a}}{\sigma_{\rm b}}\, EXP[\frac{z_{\rm a}^2 - z_{\rm b}^2}{2}]}$$

Numerical Answer Adding to the table in the question:

x_1	$\mu_{\rm a}$	$\sigma_{\rm a}$	$z_{\rm a}$	$\mu_{\rm b}$	$\sigma_{\rm b}$	$z_{\rm b}$	$\frac{\mathrm{P}[\mathrm{N_a}]}{\mathrm{P}[\mathrm{N_b}]}$	$\frac{z_{\rm b}^2 - z_{\rm a}^2}{2}$	$EXP[\frac{z_{\rm b}^2 - z_{\rm a}^2}{2}]$
4	-1	2	$^5/_2$	18	4	$^{-7}/_2$	2	$\frac{(\frac{-7}{2})^2-(\frac{5}{2})^2}{2} = \frac{\frac{49-25}{4}}{2} = \frac{\frac{24}{4}}{2} = 3$	$EXP[3] \approx 20$

$$\frac{\mathrm{P}[\mathrm{N_a}|x]}{\mathrm{P}[\mathrm{N_b}|x]} = \frac{\mathrm{P}[\mathrm{N_a}]}{\mathrm{P}[\mathrm{N_b}]}\frac{\sigma_{\rm b}}{\sigma_{\rm a}}\, EXP[\frac{z_{\rm b}^2 - z_{\rm a}^2}{2}] \approx 2 \cdot \frac{4}{2} \cdot 20 \approx 80 \quad \text{so} \quad \mathrm{P}[\mathrm{N_a}] \approx \frac{80}{81}$$

Your Name: _____

Consider a typically classification problem. For example one with m features $\mathrm{F}_1, \ldots, \mathrm{F}_m$, which may be symptoms, e.g. fever, coughing, etc.; while the class to be predicted is the person's condition (healthy or disease, etc.).

Consider the following statements:

$$\mathrm{P}[\mathrm{F}_1, \ldots, \mathrm{F}_m] \Leftrightarrow \prod_{i=1}^{m} \mathrm{P}[\mathrm{F}_i] \tag{1}$$

$$\mathrm{P}[\mathrm{F}_1, \ldots, \mathrm{F}_m|\mathrm{C}] \Leftrightarrow \mathrm{P}[\mathrm{F}_1|\mathrm{C}]\,\mathrm{P}[\mathrm{F}_2|\mathrm{C}, \mathrm{F}_1]\,\mathrm{P}[\mathrm{F}_3|\mathrm{C}, \mathrm{F}_1, \mathrm{F}_2] \cdots \mathrm{P}[\mathrm{F}_m|\mathrm{C}, \mathrm{F}_1, \ldots, \mathrm{F}_{m-1}] \tag{2}$$

$$\mathrm{P}[\mathrm{C}|\mathrm{F}_1, \ldots, \mathrm{F}_m] \Leftrightarrow \mathrm{P}[\mathrm{C}, \mathrm{F}_1]\,\mathrm{P}[\mathrm{C}, \mathrm{F}_2, \mathrm{F}_1]\,\mathrm{P}[\mathrm{C}, \mathrm{F}_3, \mathrm{F}_1, \mathrm{F}_2] \cdots \mathrm{P}[\mathrm{C}, \mathrm{F}_m, \mathrm{F}_1, \ldots, \mathrm{F}_{m-1}] \tag{3}$$

$$\mathrm{P}[\mathrm{C}|\mathrm{F}_1, \ldots, \mathrm{F}_m] \Leftrightarrow \prod_{i=1}^{m} \mathrm{P}[\mathrm{C}, \mathrm{F}_i] \tag{4}$$

$$\mathrm{P}[\mathrm{F}_1, \ldots, \mathrm{F}_m|\mathrm{C}] \Leftrightarrow \prod_{i=1}^{m} \mathrm{P}[\mathrm{F}_i|\mathrm{C}] \tag{5}$$

For each blank in the table below fill in one of {always, often, seldom, hardly}, where always means always equal, often means often approximately equal, seldom means usually unequal, and hardly means hardly ever equal.

Eq num:	(1)	(2)	(3)	(4)	(5)
Equal?					

Why? Give a the reason for your answers (可以用中文).

Solution: Solution:

Eq num:	(1)	(2)	(3)	(4)	(5)
Equal?	seldom	always	hardly	hardly	often

Reasons

(1) Assumes features are independent of each other. This is possible, but seldom happens because in a typical classification problem the features tend to depend on each other via the class. For example people with fevers are more likely to cough, because those two features both correlate with the health of the person.

(2) Is an instance of the "chain rule" for conditional probability and is identically true.

(3) Is hardly ever true. Note that in general the right hand side of equation (3) gets smaller and smaller with increasing m, but the left hand side doesn't.

(4) Is also hardly ever true. Same reason as for (3).

(5) Is often approximately true. Note that this is the approximation behind the Naïve Bayes classifier, which often works reasonably well as a classifier.

Your Name: _____

	start	unknowns	finish
example 1.	$(a-b)(a+b)$	$\equiv \underline{?}^{\,2} - \underline{?}^{\,2}$	$\Rightarrow a^2 - b^2$
example 2.	$x!$	$\equiv \Gamma(\underline{?})$	$\Rightarrow \Gamma(x+1)$
problem 1.	$\binom{n}{k}$	$\equiv \frac{\Gamma(\underline{?})}{\Gamma(\underline{?})\Gamma(\underline{?})}$	$\Rightarrow$ _____
problem 2.	Binomal$(k\|p,n)$	$\underline{?}$ Beta$(\underline{?}, \underline{?}, \underline{?})$	$\Rightarrow$ _____Beta(_____)

Problem 1. Fill in problem 1. in the table above relating $\binom{n}{k}$ to $\frac{\Gamma(\underline{?})}{\Gamma(\underline{?})\Gamma(\underline{?})}$.

The Beta distribution is defined as:

$$\text{Beta}(x|a,b) \stackrel{\text{def}}{=} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

Problem 2. Express the Beta distribution parameters (a,b) and data point x in terms of (k, p and n), such that

$$\text{Binomal}(k|p,n) = \text{F}(k,p,n)\text{Beta}(x|a,b)$$

Where $\text{F}(,p,n)$ is a simple function of k, p, and n.

Solution:

The Binomial and Beta distribution are defined as:

$$\text{Binomial}(k|p,n) \stackrel{\text{def}}{=} \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{Beta}(x|a,b) \stackrel{\text{def}}{=} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

We can see that x corresponds to p, so first rename p as x in Binomial(k | x, n).
Also consider the normalization terms:

By inspecting the form of these equations, we see likely correspondences:
$p \Leftrightarrow x,\ a \Leftrightarrow k,\ b \Leftrightarrow n-k.$

Trying this we obtain:

$$\begin{aligned}
\text{Binomial}(k|p,n) &= \binom{n}{k} p^k (1\text{-}p)^{n-k} \\
&= \binom{a+b}{a} p^a\,(1\text{-}p)^b && \text{Substituting } k \to a,\ n\text{-}k \to b\text{: } n \to a+b \\
&= \frac{(a+b)\,\Gamma(a+b)}{ab\,\Gamma(a)\,\Gamma(b)}\ p^{a-1}\,p\ \ (1\text{-}p)^{b-1}\,(1\text{-}p) \\
&= \frac{(a+b)\,\Gamma(a+b)}{ab\,\Gamma(a)\,\Gamma(b)}\ p^{a-1}(1\text{-}p)^{b-1}\ p(1\text{-}p) \\
&= \frac{(a+b)\ \text{Beta}(p|a,b)\ p(1\text{-}p)}{a\,b} && \text{Definition of Beta(p|a, b)} \\
&= p(1\text{-}p)\ \frac{n}{k\,(n\text{-}k)}\ \text{Beta}(p|k,n\text{-}k) && \text{Restoring: } a \to k,\ b \to n\text{-}k \text{ and rearranging}
\end{aligned}$$