Final examination for course of Fundamentals of Statistical Machine Learning (20190110).
This exam is closed book, closed notes.

Your Name: ____

1. Let f denote a probability density function over the real numbers, with mean f_μ, standard deviation f_σ and variance f_{σ^2}.

   Let g be the sum of n numbers indepently sampled from f. Denote its mean, standard deviation and variances as: g_μ, g_σ and g_{σ^2} respectively.

   In general, which (if any) of the following hold? Multiple answers allowed.

   A. $g_\mu = nf_\mu$ B. $g_\sigma = nf_\sigma$ C. $g_{\sigma^2} = nf_{\sigma^2}$

   Answer:______**A, C**______

   Solution: Recalling that $f_\mu \overset{\text{def}}{=} \mathrm{E}[f_i]$ and by the linearity of expectation $\mathrm{E}[\mathrm{X}+\mathrm{Y}] \equiv \mathrm{E}[\mathrm{X}] + \mathrm{E}[\mathrm{Y}]$.

   $$g_\mu = \mathrm{E}\left[\sum_{i=1}^{n} f_i\right] = \sum_{i=1}^{n} \mathrm{E}\left[f_i\right] = nf_\mu \checkmark$$

   Recalling that $f_{\sigma^2} \overset{\text{def}}{=} \mathrm{E}[(f_i - f_\mu)^2]$

   $$g_{\sigma^2} = \mathrm{E}\left[\sum_{i=1}^{n} (f_i - f_\mu)^2\right] = \sum_{i=1}^{n} \mathrm{E}\left[(f_i - f_\mu)^2\right] = nf_{\sigma^2} \checkmark$$

   On the other hand, $g_\sigma \overset{\text{def}}{=} \sqrt{g_{\sigma^2}} = \sqrt{nf_{\sigma^2}} = \sqrt{n}\sqrt{f_{\sigma^2}} = \sqrt{n}f_\sigma$.

Your Name: ____________________

2. In the year 20XX the population distribution of the continents may be approximately:

Continent	Fraction
Asia	1/2
Africa	1/4
Europe	1/8
Americas	3/32
Oceania	1/32

The information theoretic entropy of this distribution is approximately 1.85.

Here "Americas" means North and South America combined. Assuming that North and South America have equal populations, what would the information theoretic entropy be if they are treated separately (i.e. the category "Americas" is split into North and South America).

A. 1.85 B. 1.80 C. 1.65 D. 1.75 E. 1.90 F. 1.95 G. 1.60 H. 2.00

Answer: **F**

Solution: The closest value is 1.95.

This problem involves updating an entropy calculation after partitioning an element.

In general, let P = $p_1 \dots p_{j-1}, p_j, p_{j+1} \dots p_n$, and S = $s_1 \dots s_m$ be probability vectors. Imagine Q is obtained from P by splitting element p_j into m distinct elements according to S.

Q = $p_1 \dots p_{j-1},\ p_j s_1, p_j s_2 \cdots p_j s_m,\ p_{j+1} \cdots p_n$.

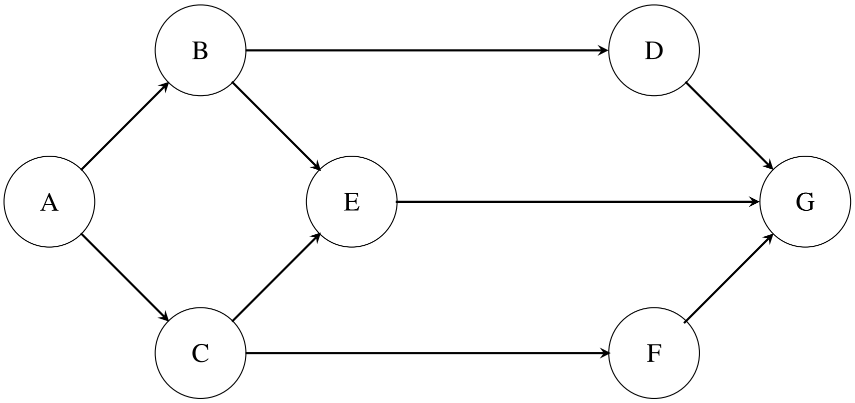
$$\mathrm{H(Q)} = \mathrm{H(P)} + p_j \mathrm{H(S)}$$

In this particular problem p_j corresponds to the probability of the Americas (3/32); and S the 50/50 splitting of the America's.

So the entropy after spliting the Americas is: $\approx 1.85 + 3/32 \cdot 1 \approx 1.85 + 0.1 = 1.95$

(More precisely, the entropy ≈ 1.94516).

Your Name: ______________________________



3.

Let $W \perp\!\!\!\perp X \mid Y,Z$ denote the statement: W and X are independent given Y and Z.

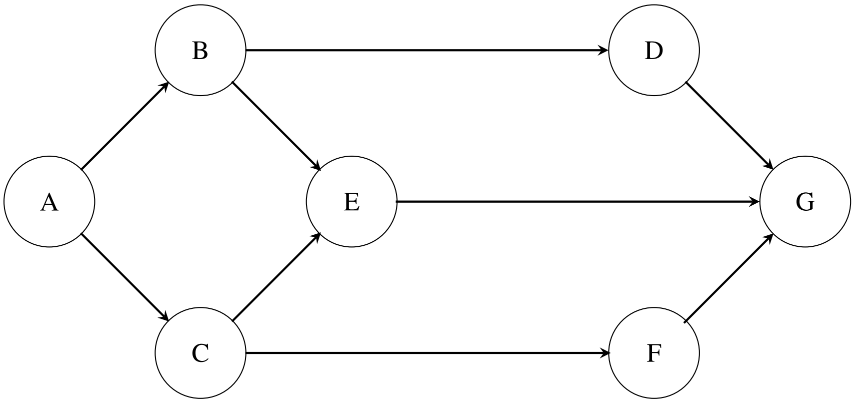
And $W \perp\!\!\!\perp X,Y \mid Z$ denote the statement: W and the set {X,Y} are independent given Z.

Which (if any) of the following choices is true? Multiple answers allowed.

1. $B \perp\!\!\!\perp F \mid A,E$ 2. $B \perp\!\!\!\perp F \mid A,E,G$ 3. $C \perp\!\!\!\perp D \mid B,F$ 4. $C \perp\!\!\!\perp D \mid B,G$ 5. $C \perp\!\!\!\perp B,D \mid A$

Answer: **3, 5**

Solution:



Notation: boldface is used to denote nodes which are conditioned on. Intuition is given based on a causal interpretation of the graph, but *in general edges in Bayes nets do not necessarily indicate causality*.

1. $B \perp\!\!\!\perp F \mid A, E$? **No**, conditionally dependent.

Reason: one interpretation of the graph is that node A is a common cause of both B and F (indirectly via C).

Formally, the path B←**A**→C→F

makes B and F dependent (**A** in boldface means it is being conditioned on).

2. $B \perp\!\!\!\perp F \mid A, E, G$? **No**, conditionally dependent; for the same reason.

3. $C \perp\!\!\!\perp D \mid B, F$? **Yes**, conditionally independent.

Again using a causal interpretation of the graph for intuition; C and D potentially depend on each other via a common cause A, and a common effect G. However, any causal effect of A on D must go through B but we already know B. Conversely, the common effect G is not a dependency because we don't know it.

More formally, the sub path A→**B**→D is d-separated because we are conditioning on B, and the colliders: C→E←**B**, E→G←D, **F**→G←D are all d-separated because we are not conditioning on E or G (nor any descendents thereof).

One can confirm by inspection that all paths from C to D must include one of those four sub-paths. ✓

4. $C \perp\!\!\!\perp D \mid B, G$? **No**, conditionally dependent.

Intuition: In this case we know the state of the common effect G.

Formally the path C→F→**G**←D makes C dependent on D.

5. $C \perp\!\!\!\perp B, D \mid A$? **Yes**, conditionally independent.

Intuitively, knowing A, the common cause of B and C, removes dependency between B and C.

Formally, sub-paths C←**A**→B, C→E←B, F→G←D, F→G←E all are d-separated and all paths from C to B or from C to D include at least one of those d-separated sub-paths.

4. The melting temperatures of magnesium (Mg) and aluminium (Al) are Mg:650 and Al:660 °C respectively.

   Suppose we will receive a sample of either Mg or Al, but do not know in advance which one we will receive. Based on past experience, we expect a 75% chance of Al (and therefore 25% of Mg).

   Further suppose we can measure melting temperature, but with measurement error. The error is unbiased (has a mean of zero) and is well approximated by a normal distribution with a standard deviation of 2°C.

   One sample is received and its melting temperature is measured to be 654 ±2°C.

   Which one of the following is a reasonable estimate of the probability that the sample is Mg?

   A. $1/3e^2$ B. $1/3\sqrt{e}$ C. $1/3e^4$ D. $1/3e^{0.4}$ E. $1/4e^2$ F. $1/4\sqrt{e}$ G. $1/4e^4$ H. $1/4e^{0.4}$ I. None of the above.

   Answer: **None of the above**

Solution: According to the description of the noise distribution, the likelihood ratio of a measurement error of $z_1 = \sigma x_1$ versus $z_2 = \sigma y$ degrees is:

$$\frac{exp(-\frac{z_1^2}{2})}{exp(-\frac{z_2^2}{2})} = \frac{exp(\frac{z_2^2}{2})}{exp(\frac{z_1^2}{2})} = e^{\frac{z_2^2 - z_1^2}{2}}$$

In this problem, the z-scores for Al and Mg are: $\frac{660-654}{2} = 3$ or $\frac{654-650}{2} = 2$ respectively.

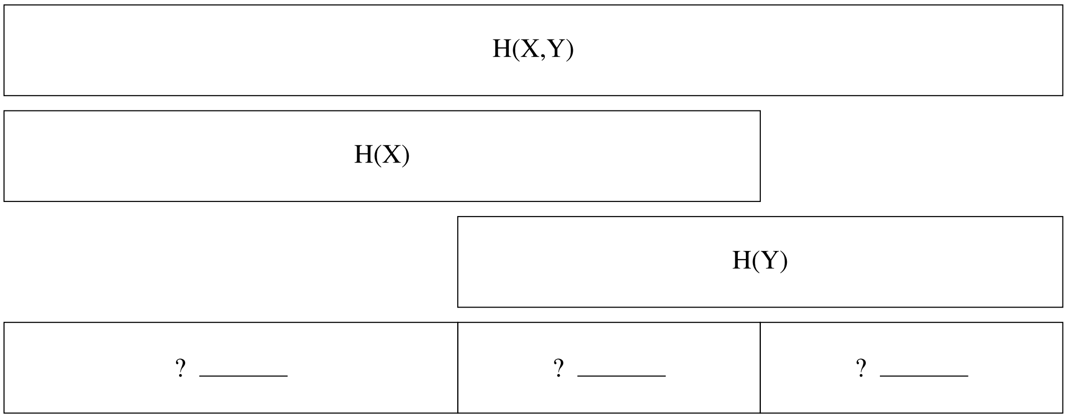
So the likelihood ratio of Mg:Al is $e^{\frac{3^2-2^2}{2}} = e^{\frac{5}{2}} = \sqrt{e^5} \approx 12.18$

Multiplying by the prior odds of 25:75=1:3, the posterior odds of Mg:Al are $\frac{1}{3}e^{\frac{5}{2}}$.

Since Al and Mg are the only two possibilities, we may use the relationship: probability $= \dfrac{\text{odds}}{1 + \text{odds}}$

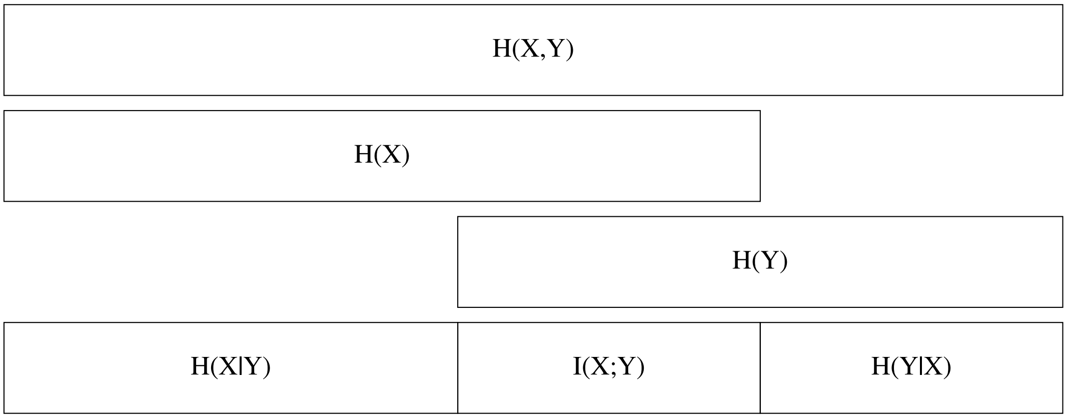
$$\text{P[Mg]} = \frac{\frac{1}{3}e^{\frac{5}{2}}}{\frac{1}{3}e^{\frac{5}{2}} + 1} = \frac{e^{\frac{5}{2}}}{e^{\frac{5}{2}} + 3} \approx 0.8024$$

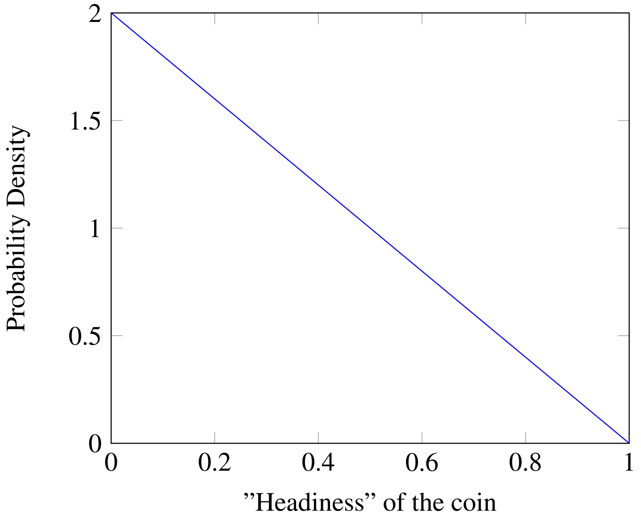
5.



In the diagram above, fill in the quantities marked by ? ______.

Solution:





6. The graph above represents the posterior belief in the “headiness” of a coin after observing one sample of data which happened to be a **tail**. The prior belief regarding the headiness of the coin had the form of a beta distribution with parameters (a, b).

   For your reference: the beta distribution Beta(r:a,b) is defined as: $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}r^{a-1}(1-r)^{b-1}$.

   Question: What values of (a, b) did the prior have?

   a:_______**1**_______ b:_______**1**_______

Solution: The prior was Beta(r:1,1). Why? By inspection of the plot above we can see that the posterior distribution is $2(1-r)$. Since the observed data is a single tail, the posterior probability is

$$Beta(r : a, b+1) = \frac{\Gamma(a+b+1)}{\Gamma(a)\Gamma(b+1)}r^{a-1}(1-r)^{b}$$

To make that look like the line $2(1-r)$, the a and b prior parameters must have both been 1. To check this substitute those values into the posterior:

$$Beta(r : 1, 1+1) = \frac{\Gamma(3)}{\Gamma(1)\Gamma(2)}r^{0}(1-r)^{1} = \frac{2!}{0!\,1!}r^{0}(1-r)^{1} = 2(1-r)\ ✓$$

The above solution uses the notion of “pseudocounts”. A more basic way to confirm that the prior was Beta(r:1,1) requires the derivation of P[D,r] and P[D]. In this case the data D is one coin flip resulting in a tail. Under a uniform prior the probability density P[D,r] is equal to P[D|r] = (1-r). By symmetry, intuitively P[D] should be ½, evenly splitting the probability of a head or a tail. More formally:

$$\mathrm{P[D]} = \int_{\rho=0}^{1} \mathrm{P[D|\rho]P[\rho]} = \int_{\rho=0}^{1}(1-\rho)d\rho = \int_{\rho=0}^{1} d\rho - \int_{\rho=0}^{1}\rho d\rho = \rho - \frac{1}{2}\rho^2\Big|_{\rho=0}^{1} = 1 - \frac{1}{2} = \frac{1}{2}$$

Combinining,

$$\mathrm{P}[r|\mathrm{D}] = \frac{\mathrm{P}[r]\mathrm{P}[\mathrm{D}|r]}{\mathrm{P[D]}} = \frac{(1-r)}{\frac{1}{2}} = 2(1-r)✓$$