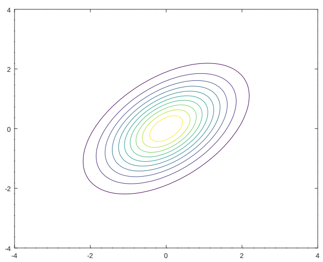
Final exam, 機器學習, Fall 2020. Closed book, no calculators/cell phones allowed. Answers may include e^2, $\sqrt{2}$, etc. but simplify when possible.

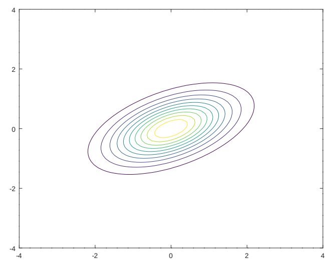
Your Name: ______________________________

Problem 1.

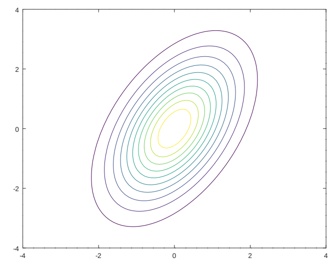
A



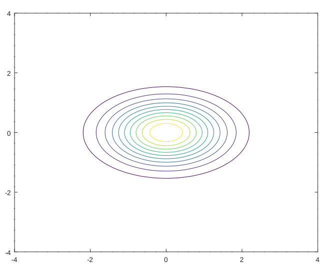
B



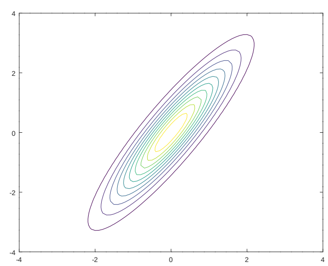
C



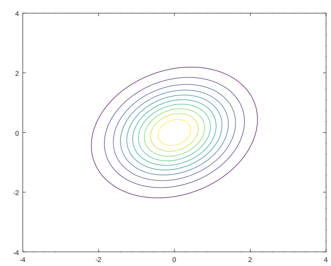
D



E



F



The above contour plots represent bivariate normal distributions $\mathcal{N}(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$, over (X,Y); with X plotted on the horizontal axis, and Y on the vertical axis. Six different plots are presented. For all six $(\mu_X, \mu_Y) = (0, 0)$ and $\sigma_x = 1$. For each distribution: $\sigma_Y \in \{0.7, 1, 1.5\}$, $\rho \in \{0.0, 0.2, 0.5, 0.9\}$.

ID	σ_Y	ρ	Comment
A	1.0	0.5	$\sigma_Y = \sigma_X$; shape is intermediate
B	0.7	0.5	σ_Y is small like in D; considering $\sigma_Y < \sigma_X$, shape is intermediate
C	1.5	0.5	σ_Y is large like in E; considering $\sigma_Y > \sigma_X$, shape is intermediate;
D	0.7	0.0	long axis parallel to x-axis so $\rho = 0$; $\sigma_Y < \sigma_X$ and smaller than σs in other plots
E	1.5	0.9	very narrow shape, so $\rho = 0.9$; σ_Y large, like in plot C
F	1.0	0.2	$\sigma_Y = \sigma_X$; shape is similar to, but not quite, a cirle

For this problem, it helps to recall that the marginal distribution of a multivariate normal is simply the parameters of the remaining (not marginalized) variables. So the relative size of σ_X and σ_Y can more or less by determined by projecting a contour onto the X and Y axes.

Your Name: ______________________________

Problem 2.

Question 2a Give (and justify) the simplest example you can find of a joint probability distribution over variables $\{A, B, C\}$. Such that A and B are pairwise independent but $A \not\perp\!\!\!\perp B \,|C$.

Solution: Many answers are possible. The classic example is A, B, and C are three boolean variables with C equal to the exclusive or of A and B, $C = A \oplus B$, so that given any two of $\{A, B, C\}$, the third is completely determined, and in particular $\mathrm{P}[A{=}1|B, C] = \begin{cases} 0 \text{ if } B = C \\ 1 \text{ otherwise} \end{cases}$

Suppose the priors of A and B are independent Bernoulli distribubions $\mathrm{P}[A]{\sim}\mathrm{Bernoulli}(0.5)$ and $\mathrm{P}[B]{\sim}\mathrm{Bernoulli}(0.5)$, where Bernoulli(0.5) is a fair coin-flip with value 0 or 1.

By this definition $\mathrm{P}[A{=}0|C{=}c] = \begin{cases} \mathrm{P}[B = 0] \text{ if } c = 0 \\ \mathrm{P}[B = 1] \text{ if } c = 1 \end{cases}$

But $\mathrm{P}[B{=}0] = \mathrm{P}[B{=}1] = 0.5$ so $\mathrm{P}[A|C]{\sim}\mathrm{Bernoulli}(0.5)$ which clearly differs from the deterministic relationship of $\mathrm{P}[A|B, C]$. Thus $A \not\perp\!\!\!\perp B \,|C$ ✓.

Question 2b Give (and justify) the simplest example you can find of a joint probability distribution over variables $\{A, B, C\}$. Such that $A \perp\!\!\!\perp B \,|C$, but A and B are **not** pairwise independent.

Solution: Many answers are possible. A simple one is if A and B are exact copies of C. These relationships are deterministic, so

$$\begin{aligned} \mathrm{P}[A{=}c|B] &= \mathrm{P}[A{=}c] = 1 \\ \mathrm{P}[A{\neq}c|B] &= \mathrm{P}[A{\neq}c] = 0 \end{aligned} \tag{1}$$

Clearly $A \perp\!\!\!\perp B \,|C$. Suppose the prior distribution of C is $P[C]{\sim}\mathrm{Bernoulli}(0.5)$. By marginalizing C out of $P[A, C]$ we obtain:

$$\begin{aligned} \mathrm{P}[A{=}0] = \; & \mathrm{P}[c{=}0]\,\mathrm{P}[A{=}0|c{=}0] \;+\; \mathrm{P}[c{=}1]\,\mathrm{P}[A{=}0|c{=}1] = \; \tfrac{1}{2}(1) + \tfrac{1}{2}(0) = \; \tfrac{1}{2} \\ \mathrm{P}[A{=}1] = \; & \mathrm{P}[c{=}0]\,\mathrm{P}[A{=}1|c{=}0] \;+\; \mathrm{P}[c{=}1]\,\mathrm{P}[A{=}1|c{=}1] = \; \tfrac{1}{2}(0) + \tfrac{1}{2}(1) = \; \tfrac{1}{2} \end{aligned}$$

So the marginal probability $\mathrm{P}[A]$ is also $\mathrm{P}[A]{\sim}\mathrm{Bernoulli}(0.5)$. The above math may be overkill though, as one could simply reason that since A is a perfect copy of C; $\mathrm{P}[C]{\sim}\mathrm{Bernoulli}(0.5)$ immediately implies $\mathrm{P}[A]{\sim}\mathrm{Bernoulli}(0.5)$.

Comparing this result to equation 1 above, $\mathrm{P}[A] \neq \mathrm{P}[A|B]$, thus A and B are not pairwise independent. ✓

Your Name: ______________________________

Problem 3.
Assume we know of two linear functions of x:

$$F_1(x) = mx + b_1; \quad F_2(x) = mx + b_2$$

with known values of m, b_1, and b_2, with $b_1 < b_2$.
Further suppose we have n points of data in the form of x, y points (e.g. the point (x=0,y=0) or (x=2,y=3), etc.) where some of the points were generated by: $y_i = F_1(x_i) + \mathcal{N}(0, \sigma_1^2)$ and some of the points were generated by $y_i = F_2(x_i) + \mathcal{N}(0, \sigma_2^2)$. We are not told which points are from which function, but we are told that the ratio of points from F_1 to those from F_2 is $\sigma_1 : \sigma_2$, i.e. the number of points from F_1 is $\frac{n\sigma_1}{\sigma_1+\sigma_2}$.

Question: in terms of parameters given above (m, b_1, b_2, σ_1, σ_2) give an optimal decision rule for classifying a point (x, y) as belonging to F_1 or F_2. Where optimal means fewest expected mistakes.

Solution: Define $d = y - mx$. Note that according to the problem formulation above:

$$\mathrm{P}[y|F_1, x] = \mathrm{P}[d|F_1] = \frac{1}{\sigma_1 \exp(\frac{(d-b_1)^2}{2\sigma_1^2})}$$

and similarly for F_2.
At the decision boundary, $P[F|x, y]$ should be the same ($= 0.5$) for F_1 and F_2.
So we should solve for:

$$1 = \frac{\mathrm{P}[F_1|x,y]}{\mathrm{P}[F_2|x,y]} = \frac{\mathrm{P}[F_1]\,\mathrm{P}[x,y|F_1]\,\frac{1}{\mathrm{P}[x,y]}}{\mathrm{P}[F_2]\,\mathrm{P}[x,y|F_2]\,\frac{1}{\mathrm{P}[x,y]}} = \frac{\mathrm{P}[F_1]\,\mathrm{P}[x,y|F_1]}{\mathrm{P}[F_2]\,\mathrm{P}[x,y|F_2]} = \frac{\sigma_1\,\mathrm{P}[x,y|F_1]}{\sigma_2\,\mathrm{P}[x,y|F_2]}$$

$$= \frac{\sigma_1\,\mathrm{P}[x]\,\mathrm{P}[y|F_1,x]}{\sigma_2\,\mathrm{P}[x]\,\mathrm{P}[y|F_2,x]} \frac{\sigma_1 \frac{1}{\sigma_1 \exp(\frac{(d-b_1)^2}{2\sigma_1^2})}}{\sigma_2 \frac{1}{\sigma_2 \exp(\frac{(d-b_2)^2}{2\sigma_2^2})}} = \frac{\exp(\frac{(d-b_2)^2}{2\sigma_2^2})}{\exp(\frac{(d-b_1)^2}{2\sigma_1^2})}$$

$$\implies \frac{(d-b_1)^2}{\sigma_1^2} = \frac{(d-b_2)^2}{\sigma_2^2} \implies \frac{|d-b_1|}{\sigma_1} = \frac{|d-b_2|}{\sigma_2} \implies \sigma_2|d-b_1| = \sigma_1|d-b_2|$$

The decision rule is predict F_1 if $\sigma_2|d-b_1| < \sigma_1|d-b_2|$, otherwise predict F_2. If $b_1 \leqq d \leqq b_2$, that rule corresponds to predict F_1 if

$$\sigma_2|d-b_1| < \sigma_1|d-b_2| \implies \sigma_2(d-b_1) < \sigma_1(b_2-d) \implies (\sigma_1+\sigma_2)d < \sigma_1 b_2 + \sigma_2 b_1$$

$$\implies d < \frac{\sigma_1 b_2 + \sigma_2 b_1}{\sigma_1+\sigma_2} = b_1 + \frac{\sigma_1(b_2-b_1)}{\sigma_1+\sigma_2}$$

Note that if $\sigma_1 \neq \sigma_2$, then there will be another decision boundary. For example if $\sigma_1 > \sigma_2$, there will be a point $d > b_2$ such that

$$\sigma_1(d-b_2) = \sigma_2(d-b_1) \implies d(\sigma_1-\sigma_2) = \sigma_1 b_2 - \sigma_2 b_1 \implies d = \frac{\sigma_1 b_2 - \sigma_2 b_1}{\sigma_1-\sigma_2}$$

$$\implies \frac{\sigma_1 b_2 - \sigma_2(b_2 + (b_1 - b_2))}{\sigma_1-\sigma_2} \implies d = b_2 + \frac{\sigma_2(b_2-b_1)}{\sigma_1-\sigma_2}$$

So when $\sigma_1 > \sigma_2$, the decision rule is: Predict F_2 if $b_1 + \frac{\sigma_1(b_2-b_1)}{\sigma_1+\sigma_2} < d < b_2 + \frac{\sigma_2(b_2-b_1)}{\sigma_1-\sigma_2}$, otherwise predict F_1 (of course $<$ can be replaced with $\leqq$, as this only affects the prediction for when the probability of F_1 versus F_2 is 50%-50%).
By symmetry, when $\sigma_1 < \sigma_2$, Predict F_1 if $b_1 - \frac{\sigma_1(b_2-b_1)}{\sigma_2-\sigma_1} < d < b_1 + \frac{\sigma_1(b_2-b_1)}{\sigma_1+\sigma_2}$ otherwise predict F_2.

Your Name: ______________________________

Problem 4.
Background:
Recall two methods we discussed for deciding priors; Laplace and Jeffreys. The Laplace method places a uniform distribution over the parameter to be estimated, while the more complicated Jeffreys method guarantees equivalent priors regardless of the problem parameterization.
The most common way to parameterize a 'coin-flipping' problem uses p: the probability of 'success' (e.g. the probability of heads for a coin). For this purposes of this question, I call this the "p-parameterization". The likelihood function is:

$$\mathcal{L}(p; n_0, n_1) = \binom{n}{n_0} (1-p)^{n_0} p^{n_1} \tag{2}$$

Where $n = n_0 + n_1$ is the total number of data samples, and n_0 and n_1 denote the number of failures and successes respectively.
We can use a beta distribution to represent the prior probability distribution of p; convenient because it is conjugate to the likelihood function. Recall the standard beta distribution is defined as:

$$\mathrm{Beta}(p; \alpha, \beta) \stackrel{\text{def}}{=} \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}, \qquad \text{where } \mathrm{B}(\alpha, \beta) \stackrel{\text{def}}{=} \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Question 4**a** Under the Jeffreys prior, what is the prior probability of $p = 0.5$ divided by that of $p = 0.75$? In other words, using the notation $\mathrm{pd}(p = x)$ to represent the probability density of $p = x$ for some $x, 0 \leqq x \leqq 1$, what is $\mathrm{pd}(p = 0.5)/\mathrm{pd}(p = 0.75)$?

Solution: Under the p-parameterization, the Jeffrey's prior is

$$\mathrm{Beta}(p; 0.5, 0.5) \stackrel{\text{def}}{=} \frac{p^{0.5-1}(1-p)^{0.5-1}}{\mathrm{B}(0.5, 0.5)} \propto \frac{1}{\sqrt{p(1-p)}}$$

$$\frac{\mathrm{pd}(p = \frac{1}{2})}{\mathrm{pd}(p = \frac{3}{4})} = \frac{\sqrt{\frac{3}{4}(1-\frac{3}{4})}}{\sqrt{\frac{1}{2}(1-\frac{1}{2})}} = \frac{\sqrt{\frac{3}{16}}}{\sqrt{\frac{1}{4}}} = \frac{\frac{\sqrt{3}}{4}}{\frac{1}{2}} = \frac{\sqrt{3}}{2}$$

Question continued on next page.

Problem 4. (continued)
An alternative parameterization of uses the ratio of the probability of success to failure: $r = \dfrac{p}{1-p}$. Here I will denote this as the "r-parameterization".

Question 4**b** Write the likelihood function in terms of r.

Solution:

$$r = \frac{p}{1-p} \implies (1-p)r = p \implies r = p + rp \implies \quad p = \frac{r}{1+r},\ (1-p) = \frac{1}{1+r}$$

Substituting these into equation 2 we obtain

$$\mathcal{L}(n_0, n_1; r) = \binom{n}{n_0} \left(\frac{1}{1+r}\right)^{n_0} \left(\frac{r}{1+r}\right)^{n_1} = \binom{n}{n_0} \frac{r^{n_1}}{(1+r)^n}$$

Question 4**c** Assuming we use Jeffreys method to compute the prior for the r-parameterization. What should $\mathrm{pd}(r=1)/\mathrm{pd}(r=3)$ be?

Solution: Since $r = 1$ and $r = 3$, correspond to $p = 0.5$ and $p = 0.75$ respectively, it is tempting to say the answer should be the same as $\frac{\mathrm{pd}(p=0.5)}{\mathrm{pd}(p=0.75)} = \frac{\sqrt{3}}{2}$. However we need to take into consideration the non-linear change of variables from p to r. Remember that while the range of r is $r \in (0, \infty)$, half of the range of $p \in [0, 0.5]$ is packed into $r \in [0, 1]$. Informally, p is densely packed into small values r, but sparsely packed for large values of r.
More precisely, since $r = \frac{p}{1-p} \implies p = \frac{r}{r+1} = 1 - \frac{1}{r+1}$

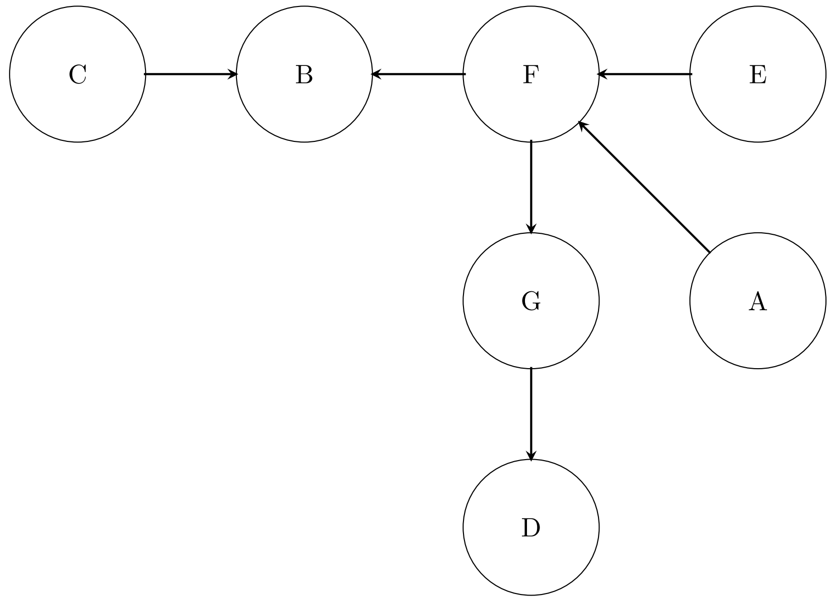
$$\frac{\partial p}{\partial r} = \frac{\mathrm{d}}{\mathrm{d}r}\left(1 - \frac{1}{r+1}\right) = \frac{1}{(r+1)^2}$$

Combining,

$$\frac{\mathrm{pd}(r=1)}{\mathrm{pd}(r=3)} = \frac{\mathrm{pd}(p=0.5)}{\mathrm{pd}(p=0.75)} \frac{\frac{\partial p}{\partial r}\Big|_{r=1}}{\frac{\partial p}{\partial r}\Big|_{r=3}} = \frac{\sqrt{3}}{2} \frac{(3+1)^2}{(1+1)^2} = \frac{\sqrt{3}}{2} \frac{4^2}{2^2} = 2\sqrt{3}$$

Your Name: ______________________________

Problem 5.



The graph above is a Bayesian network with nodes {A,B,C,D,E,F,G}, but, except A, the node labels are hidden.
The graph structure implies the following relationships:

Pairwise dependencies: A,B; A,D; A,G; B,E; D,E

Conditional independencies: A,B|F; A,D|F; A,D|G; D,F|G; D,E|F

Conditional dependencies: A,B|C; A,B|D; A,E|F; C,D|B

(at least, the above list not complete).

Question: What labeling of the nodes is consistent with those independence relationships?
In the graph at top, fill in node names.

Solution: One solution is shown above. Some hints regarding how to solve it. Notice the graph is a tree, so each pair of node has only one path joining it and most of these are simple chains. Pairwise dependencies are on the same chain, conditional independencies are from breaking the chain, so for example from ($A \not\perp\!\!\!\perp B$), ($A \not\perp\!\!\!\perp D$) we can deduce that A,B and A,D are on the same chain, while ($A \perp\!\!\!\perp B \mid F$), ($A \perp\!\!\!\perp D \mid F$) implies F blocks chains $A \rightarrow \cdots F \rightarrow \cdots \rightarrow B$, $A \rightarrow \cdots F \rightarrow \cdots D$, so F should be placed somewhere upstream of A, with B and D further upstream. Conditional dependencies, on the other hand, can be parents (or ancestors) conditioned on a child (descendent) in a "collider" structure ○→○←○. A simple example is ($A \not\perp\!\!\!\perp E|F$), where A and E are parents of F. A more complicated one is ($C \not\perp\!\!\!\perp D \mid B$), which (when given the answer at least) can be understood by first noticing (($C \not\perp\!\!\!\perp F \mid B$)) since C and F are parents of B, and then realizing that, as a descendent of F, D holds information about F, so in general ($C \not\perp\!\!\!\perp D \mid B$).