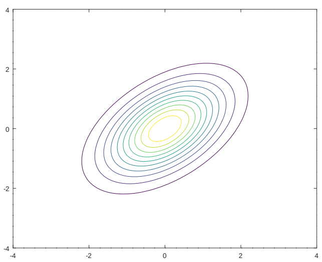
Final exam, 機器學習, Fall 2020. Closed book, no calculators/cell phones allowed. Answers may include e^2, $\sqrt{2}$, etc. but simplify when possible.

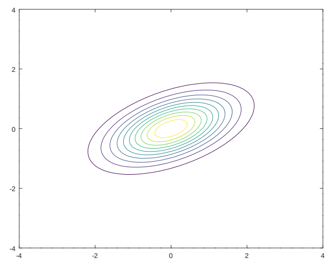
Your Name: ______________________________

Problem 1.

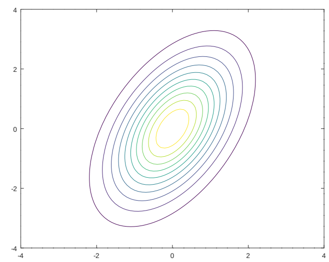
A



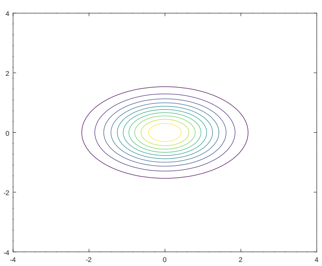
B



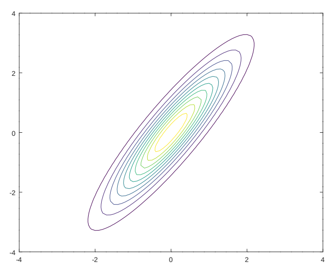
C



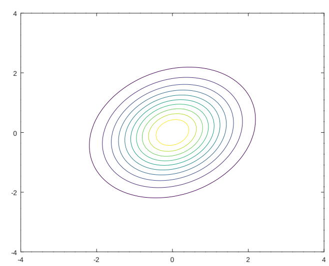
D



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			
B			
C			
D			
E			
F			

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 $A \not\perp\!\!\!\perp B \,|C$.

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.

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.

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:

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

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 $\text{pd}(p = x)$ to represent the probability density of $p = x$ for some $x, 0 \leqq x \leqq 1$, what is $\text{pd}(p = 0.5)/\text{pd}(p = 0.75)$?

Question continued on next page.

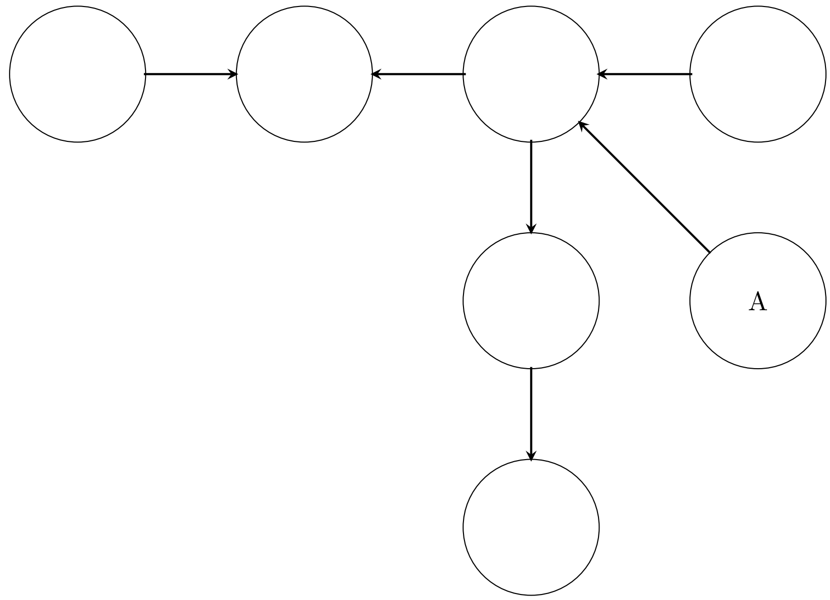
Problem 4. (continued)
An alternative parameterization of uses the ratio of the probability of success to failure: $r = \dfrac{p}{1-p}$. Here I will denote this as the "r-parameterization".

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

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?

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.