ADDITIONAL PRACTICE QUESTIONS
STA414H1S / STA2104H1S
April 2017 - revised

1. Consider revisiting any quiz-type questions asked of you on slides or otherwise during lecture. Review the three assignments and midterm.

2. Extra questions are often at the back of chapters in the commonly mentioned textbooks for this course. You might like problems here: `http://www.utstat.utoronto.ca/~radford/sta414` specifically Question 5 on Problem Set 2, as well as Questions 1, 3 and 4 in Problem Set 3.

3.

Recall that the definition of an exponential family model is:

$$f(x|\eta) = h(x)g(\eta)\exp(\eta^\top T(x))$$

where:

η	are the parameters
$T(x)$	are the sufficient statistics
$h(x)$	is the base measure
$g(\eta)$	is the normalizing constant

Consider the univariate Gaussian, with mean μ and precision $\lambda = 1/\sigma^2$ is:

$$p(D|\mu,\lambda = \prod_{i=1}^{N}\left(\frac{\lambda}{2\pi}\right)^{1/2}\exp(-\frac{\lambda}{2}(x_i-\mu)^2)$$

What are η and $T(x)$ for this distribution when it is represented in exponential family form?
Solution:

$$p(D|\mu,\lambda = (2\pi)^{-N/2}[\lambda^{1/2}\exp(-\frac{\lambda}{2}\mu^2)]^N\exp[\mu\lambda\sum_i x_i - \lambda/2\sum_i x_i^2]$$

$\eta = [\mu\lambda \;\; ; \;\; -\lambda/2]$
$T(x) = [\sum_i x_i \;\; ; \;\; \sum_i x_i^2]$

Misc questions:

1. When is the Kullback-Leibler (KL) divergence $KL(q||p) = \sum_x p(x) \log \left(\frac{q(x)}{p(x)}\right)$ zero?
2. What is the difference between the Maximum-likelihood Estimator (MLE) and the Maximum A Posteriori (MAP) estimator? When do they give the same answer?
3. What is the time complexity of evaluating the pdf of a multivariate Gaussian, given the mean and precision matrix?
4. Which of the following is *not* a valid interpretation of the objective function of variational inference?
   (a) It minimizes the KL divergence between the approximating distribution and the posterior distribution.
   (b) It maximizes a lower bound on the log marginal likelihood of the data.
   (c) It minimizes the expected negative log likelihood while staying "close" to the true prior.
   (d) None of the above.

Answers:

1. Only when $p = q$
2. MAP incorporates a prior term $p(\theta)$. They are equivalent only when $p(\theta)$ is a constant.
3. $\mathcal{O}(D^2)$
4. d, all are valid.

Answers to questions on upcoming pages:

1 (a) Noting that the three points are collinear and have zero mean, and recalling that the first principal component will occur in the direction of maximal variation and will have unity magnitude, the vector can be derived from the difference of any two points. Acceptable answers would be $(1/\sqrt{2}, 1/\sqrt{2})$ or $(-1/\sqrt{2}, -1/\sqrt{2})$.

Alternatively, you can calculate the eigenvalues of $X'X = S$ which are 0 and 4. The eigenvector corresponding to the larger eigenvalue (4) works out to be (1,1). Normalizing leads to $(1/\sqrt{2}, 1/\sqrt{2})$ as before.

(b) Three iterations are needed:
1. The -4 cluster has points $\{-3\}$ and the 0 cluster has points $\{-1, 2, 4\}$.
2. The -3 cluster has points $\{-3, -1\}$ and the $+5/3$ cluster has points $\{2, 4\}$.
3. The -2 cluster has points $\{-3, -1\}$ and the $+3$ cluster has points $\{2, 4\}$.
As there are no further changes to make to the cluster assignments, the algorithm converged.

(d)

$$P(c = 0|x) = \frac{p(x|c = 0)p(c)}{p(x)} = \frac{40(0.6)}{40(0.6) + 120(0.4)} = \frac{1}{3}$$

and $P(c = 1|x) = \frac{2}{3}$. Thus the more probable class is 1.

2. (a)

$$\frac{0.2^{k-1}(0.32)}{0.2^{k-1}(0.32) + 0.75^{k-1}(0.15)}$$

(b) 0.594
(c) 0.727

3. (a) $P(A)P(B)P(C|A, B)P(D|B)P(F|C, D)P(E|C)$
(b) T, F, F, F, F, T

1. (10 marks) Short answer questions.

a) Consider a 2D dataset with 3 examples: $(-1,-1),(0,0),(1,1)$. If you apply PCA, what will be the first principal component? Provide it as a 2D vector. (2 marks)

b) Consider a 1D dataset with 4 examples: $-3,-1,2,4$. By hand, apply k-means clustering until convergence, assuming that the initial prototypes are -4 and 0. For each iteration, report the assignments of examples to prototypes and the new values of the prototypes. (3 marks)

d) Suppose you use naive Bayes to classify a 3D test case as being from class 0 or class 1. For class 0, the likelihoods of the 3 inputs are 2, 5 and 4, whereas for class 1, the likelihoods are 2, 3 and 20. The prior probabilities are 0.6 and 0.4 for class 0 and class 1. What are the posterior probabilities and what is the most probable class? (3 marks)

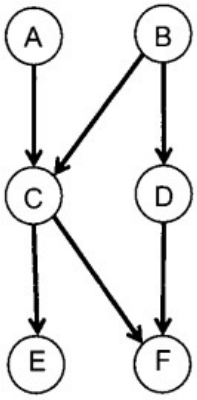
2. (10 marks) Bayesian inference. In a simplified model of driving tests, there are two types of drivers: bad drivers and good drivers. For a single driving test, a good driver has a 80% probability of passing and a bad driver has a 25% probability of passing. The prior probability that a driver is good is 0.4.

a) If a driver passes the test on the k-th attempt, what is the probability that he or she is a good driver? (3 marks)

b) Suppose we know that a driver passed the test on either the first or second attempt, $k \leq 2$. What is the probability that he or she is a good driver? (3 marks)

c) Suppose we know that a driver passed the test on either the first or second attempt, $k \leq 2$. What is the probability that he or she succeed on the first try, $k = 1$? (4 marks)

3. (10 marks) Graphical models: Properties and conditional independence. Consider the following BN.



a) Write down the formula of the joint probability distribution represented by the BN in terms of conditional and marginal probabilities. (2 marks)

b) Circle T or F for each of the following conditional independence statements of the random variables in the Bayesian network above. (3 marks)

$C \perp\!\!\!\perp D \mid B$	T / F
$C \perp\!\!\!\perp D \mid B, F$	T / F
$B \perp\!\!\!\perp F \mid D$	T / F
$A \perp\!\!\!\perp D \mid C$	T / F
$A \perp\!\!\!\perp F \mid C$	T / F
$A \perp\!\!\!\perp F \mid B, C$	T / F