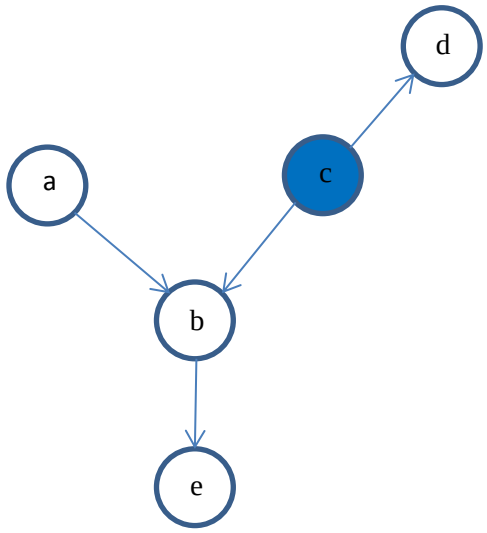
CSCI 59000: Machine Learning
Assignment 4
Instructor: Murat Dundar
Due: 4/27/2018

1. (20 points). Refer to the directed graph below for parts a through f. Shaded nodes indicate observed nodes. Suppose all the conditional probability tables associated with this graph is given.

a) (4 points) Write p(a,b,c,d,e) in terms of the conditional and prior probabilities of each node.
b) (3 points) (True,False) a is independent of d
c) (3 points) (True,False) d is independent of e
d) (5 points) Please describe how to use the sum-product algorithm to compute p(e).
e) (5 points) Please describe how to use the max-sum algorithm to compute argmax p(a,b,c,d,e).



See next page.

2. (20 points) Given a training set $\{x_i\}_{i=1}^N$ the smallest hyper-sphere fitting this data can be obtained by solving the following primal problem with respect to R and *a* which is equivalent to solving the dual problem with respect to α, i.e. KKT multipliers.

Primal:

$$\min_{(R,a)\in R^{d+1}} \quad R^2$$

$$\text{s.t.} \quad R^2 \geq (x_i - a)^T(x_i - a)\ \forall i$$

Dual:

$$\min_{\alpha\in R^N} \quad \sum_{i=1}^N \alpha_i x_i^T x_i + \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j x_i^T x_j$$

$$\text{s.t.} \quad \sum_{i=1}^N \alpha_i = 1$$

(a) (8 points) Suppose that the training data contains outliers and thus, forcing each sample to remain inside the hyper-sphere unjustifiably increase the radius of the smallest fitting hyper-sphere. Instead of finding the smallest hyper-sphere fitting all data we want to find the best fitting hyper-sphere which will allow for some of the samples to remain outside the hypersphere. Make the necessary modifications in the primal problem such that optimization with respect to R and *a* yields the best fitting hyper-sphere.

(b) (8 points) Write the dual form of the problem in part a starting with the dual form written above and derive an expression for the center, *a*, of the hyper-sphere in terms of KKT multipliers. Explain how the expression for *a* for best fitting hyper-sphere changes compared to the smallest fitting hyper-sphere.

(c) (4 points) Fitting a hyper-sphere onto the data can sometimes be quite restrictive. Therefore it would make more sense to fit a hyper-sphere in some unknown feature space which may generate arbitrarily defined shapes in the input space when appropriate kernel functions are selected. Explain how you can take advantage of the kernel trick to compute the distance of a point from the center *a* of the best fitting hyper-sphere in the feature space.

3. (20 points) Consider a special case of a Gaussian mixture model in which the covariance matrices Σ_k of the components are all constrained to have a common value σI, i.e., a common spherical matrix. Derive the EM equation for σ for maximizing the likelihood function under such a model.

4. (20 points) Solve question 11.13 from the PRML textbook and implement a Gibbs sampler to obtain samples from p(μ,τ). Iterate the sampler at least for 10,000 times and plot the values of μ and τ. How long do you think the burn-in period of the sampler was? Use μ_0=0, s_0=1, a=1, b=1.

See next page.

5. (20 points)

a. For an M node, K state directed Markov chain prove that one needs K -1 + (M-1)K(K-1) many parameters to model the joint distribution.

b. In a directed graph prove that any node is independent of the rest of the graph given the Markov blanket of that node.