PCML 2014: Sample Exam Solutions

Mohammad Emtiyaz Khan
EPFL

January 7, 2015

Abstract

I give solutions to the sample exam. Answers are indicated with **A**. Some important things to note are indicated with **Note**. This also gives you an idea of how to write your answers in the final exam to get full marks. The key is to show all steps of your derivations clearly (even though you do not arrive at the correct answer).

Weighted least-squares : (a) Derive the normal equations for this cost function ...

A. Define $\mathbf{W} := \text{diag}([w_1, w_2, \ldots, w_N])$. Taking derivative of the cost function, we get,

$$\partial \mathcal{L}\boldsymbol{\beta} = \sum_n w_n(y_n - \boldsymbol{\beta}^T\widetilde{\mathbf{x}}_n)\widetilde{\mathbf{x}}_n \tag{1}$$

$$= -\sum_n w_n y_n \widetilde{\mathbf{x}}_n + \sum_n w_n \boldsymbol{\beta}^T\widetilde{\mathbf{x}}_n\widetilde{\mathbf{x}}_n \tag{2}$$

$$= -\sum_n w_n y_n \widetilde{\mathbf{x}}_n + \sum_n w_n \widetilde{\mathbf{x}}_n\widetilde{\mathbf{x}}_n^T\boldsymbol{\beta} \tag{3}$$

$$= -\sum_n w_n y_n \widetilde{\mathbf{x}}_n + \left(\sum_n w_n \widetilde{\mathbf{x}}_n\widetilde{\mathbf{x}}_n^T\right)\boldsymbol{\beta} \tag{4}$$

$$= -\widetilde{\mathbf{X}}^T\mathbf{W}\mathbf{y} + \widetilde{\mathbf{X}}^T\mathbf{W}\widetilde{\mathbf{X}}\boldsymbol{\beta} \tag{5}$$

The normal equation therefore is $\widetilde{\mathbf{X}}^T\mathbf{W}(\mathbf{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta}) = 0$.

(b) Discuss the conditions under which the solution $\boldsymbol{\beta}^*$ is unique.

A. For a unique solution, we require $\widetilde{\mathbf{X}}^T\mathbf{W}\widetilde{\mathbf{X}}$ to be invertible. Since $\mathbf{W}$ is positive definite, the matrix will be invertible if $\widetilde{\mathbf{X}}$ is full column rank.

(c) Assuming that these conditions hold, write down the expression for the unique solution.

A. $\boldsymbol{\beta}^* = (\widetilde{\mathbf{X}}^T\mathbf{W}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\mathbf{W}\mathbf{y}$.

(d) Derive a probabilistic model ....

A. It is easy to see that the weights w_n can be treated as the inverse of Gaussian variances. Therefore the probabilistic model would be the following:

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \prod_{n=1}^{N} \mathcal{N}(y_n|\boldsymbol{\beta}^T\widetilde{\mathbf{x}}_n, 1/w_n) \tag{6}$$

Multi-class classification: Following the derivation of logistic regression,

(a) Derive the log-likelihood for this model.

A. Define the vector $\widetilde{\mathbf{y}}_n$ such that $\widetilde{y}_{nk} = 1$ when $y_n = k$ and rest of entries of $\widetilde{\mathbf{y}}_n$ are zero.

$$\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \log \prod_{n=1}^{N} p(y_n|\mathbf{x}_n, \boldsymbol{\beta}) \tag{7}$$

$$= \log \prod_{n:y_n=1} p(y_n = 1|\mathbf{x}_n, \boldsymbol{\beta}) \prod_{n:y_n=2} p(y_n = 2|\mathbf{x}_n, \boldsymbol{\beta}) \dots \prod_{n:y_n=K} p(y_n = K|\mathbf{x}_n, \boldsymbol{\beta}) \tag{8}$$

$$= \log \prod_{k=1}^{K} \prod_{n=1}^{N} [p(y_n = k|\mathbf{x}_n, \boldsymbol{\beta})]^{\widetilde{y}_{nk}} \tag{9}$$

$$= \sum_{k=1}^{K} \sum_{n=1}^{N} \widetilde{y}_{nk} \log p(y_n = k|\mathbf{x}_n, \boldsymbol{\beta}) \tag{10}$$

$$= \sum_{k=1}^{K} \sum_{n=1}^{N} \widetilde{y}_{nk} \left[\eta_{nk} - \log \sum_{j=1}^{K} \exp(\eta_{nj}) \right] \tag{11}$$

$$= \sum_{k=1}^{K} \sum_{n=1}^{N} \widetilde{y}_{nk} \left[\boldsymbol{\beta}_k^T \widetilde{\mathbf{x}}_n - \log \sum_{j=1}^{K} \exp(\boldsymbol{\beta}_j^T \widetilde{\mathbf{x}}_n) \right] \tag{12}$$

$$= \sum_{k=1}^{K} \sum_{n=1}^{N} \widetilde{y}_{nk} \boldsymbol{\beta}_k^T \widetilde{\mathbf{x}}_n - \sum_{n=1}^{N} \log \sum_{j=1}^{K} \exp(\boldsymbol{\beta}_j^T \widetilde{\mathbf{x}}_n) \tag{13}$$

Last step is obtained since $\sum_k y_{nk} = 1$. Notice the similarity to logistic regression.

Note: You get full marks even if you skip the last step. The last step is useful for the next part.

(c) Derive the gradient with respect to $\boldsymbol{\beta}_k$.

A. Taking the derivative wrt $\boldsymbol{\beta}_k$, we get,

$$\sum_{n=1}^{N} \widetilde{y}_{nk} \widetilde{\mathbf{x}}_n - \sum_{n=1}^{N} \frac{\exp(\boldsymbol{\beta}_k^T \widetilde{\mathbf{x}}_n)}{\sum_{j=1}^{K} \exp(\boldsymbol{\beta}_j^T \widetilde{\mathbf{x}}_n)} \widetilde{\mathbf{x}}_n \tag{14}$$

Note: You will get full marks if you write the above expression. Note the similarity with the logistic regression. We can write the normal equation in the same form as logistic regression.

$$\widetilde{\mathbf{X}}^T [\widetilde{\mathbf{y}}_k - \mathcal{S}(\widetilde{\mathbf{X}} \beta_k)] = 0 \tag{15}$$

where $\widetilde{\mathbf{y}}_k$ is the vector containing $\widetilde{y}_{nk}$ for all n, $\mathcal{S}(\boldsymbol{\eta})$ is the softmax function which is an extension of the σ function to multi-class.

(b) Show that the negative of the log-likelihood is convex.

A. Negative of the log-likelihood is the following:

$$-\sum_{k=1}^{K} \sum_{n=1}^{N} \widetilde{y}_{nk} \boldsymbol{\beta}_k^T \widetilde{\mathbf{x}}_n + \sum_{n=1}^{N} \log \sum_{j=1}^{K} \exp(\boldsymbol{\beta}_j^T \widetilde{\mathbf{x}}_n) \tag{16}$$

Since we know that sum of convex functions is convex, it is ok to ignore the sum over n and prove that the following is convex:

$$-\sum_{k=1}^{K} \widetilde{y}_{nk} \boldsymbol{\beta}_k^T \widetilde{\mathbf{x}}_n + \log \sum_{j=1}^{K} \exp(\boldsymbol{\beta}_j^T \widetilde{\mathbf{x}}_n) \tag{17}$$

First term is linear in $\boldsymbol{\beta}_k$, so we only need to prove that the second term is convex which will be true if we prove that log-sum-exp $\log(e^{t_1} + e^{t_2} + \ldots + e^{t_K})$ is a convex function. This was given as an exercise for online assignment so you should know how to prove this. A straightforward way to proved this is to show that the second derivative is positive-definite, although there are other proofs.

Note: I don't expect you to prove convexity of log-sum-exp, rather you should know it as a fact that it is true. I gave it as an assignment question for that reason.

Poisson regression : Following the derivation of logistic regression,

(a) Derive the log-likelihood for this model.

$$\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \sum_n \log p(y_n|\mathbf{x}_n, \boldsymbol{\beta}) = \sum_n \log \frac{e^{y_n \eta_n}}{y_n!} e^{-e^{\eta_n}} = \sum_n y_n \mathbf{x}_n^T \boldsymbol{\beta} - \exp\left(\mathbf{x}_n^T \boldsymbol{\beta}\right) + \text{cnst} \tag{18}$$

$$= \mathbf{y}_n^T \widetilde{\mathbf{X}} \boldsymbol{\beta} - \sum_n \exp\left(\mathbf{x}_n^T \boldsymbol{\beta}\right) + \text{cnst} \tag{19}$$

Note: The last step is not necessary to get full marks but it helps for the next part.

(b) Derive the normal equations.

A. Taking the gradient wrt $\boldsymbol{\beta}$,

$$\mathbf{g} = \widetilde{\mathbf{X}}^T \mathbf{y}_n - \sum_n \exp\left(\mathbf{x}_n^T \boldsymbol{\beta}\right) \mathbf{x}_n = \widetilde{\mathbf{X}}^T (\mathbf{y}_n - \exp\left(\widetilde{\mathbf{X}}^T \boldsymbol{\beta}\right)) = 0 \tag{20}$$

Note: In the second step, I have abused the notation. The term $\exp\left(\widetilde{\mathbf{X}}_n^T \boldsymbol{\beta}\right)$ producess a vector. The first step itself is enough to get full marks. I wrote the second step to show that the normal equation takes very similar form to that of logistic regression.

(c) Derive the Hessian.

A. Taking the gradient wrt $\boldsymbol{\beta}$,

$$\mathbf{H} := -\sum_n \exp\left(\widetilde{\mathbf{X}}^T \boldsymbol{\beta}\right) \mathbf{x}_n \mathbf{x}_n^T = -\widetilde{\mathbf{X}}^T \mathbf{S} \widetilde{\mathbf{X}} \tag{21}$$

where $\mathbf{S}$ is a diagonal matrix containing $\exp\left(\widetilde{\mathbf{X}}^T \boldsymbol{\beta}\right)$ as the diagonal.

Note: Again the derivation is almost identical to logistic regression.

(d) Is the negative of log-likelihood convex? Prove your answer.

A. $\mathbf{S}$ is positive definite, thereofre the negative of Hessian is positive definite when $\widetilde{\mathbf{X}}$ is full column rank. Therefore the function is convex.

(e) Write down the Newton's update and discuss its complexity.

A. The Newton's update in the k'th iteration is shown below.

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - \mathbf{H}_k^{-1} \mathbf{g}_k \tag{22}$$

$$\mathbf{H}_k := -\widetilde{\mathbf{X}}^T \text{diag}\left[\exp\left(\widetilde{\mathbf{X}}^T \boldsymbol{\beta}_k\right)\right] \widetilde{\mathbf{X}} \tag{23}$$

$$\mathbf{g}_k := \widetilde{\mathbf{X}}^T \left[\mathbf{y}_n - \exp\left(\widetilde{\mathbf{X}}^T \boldsymbol{\beta}_k\right)\right] \tag{24}$$

Computation complexity is $O(ND^2 + D^3)$, same as logistic regression.

EM for mixture of Bernoulli Answer the following questions.

1. Rewrite the likelihood $p(\mathbf{x}_n|\boldsymbol{\theta}, \mathbf{r})$ in terms of r_{nk}.

A. $p(\mathbf{x}_n|\boldsymbol{\theta}, \mathbf{r}) = \prod_{k=1}^{K} \left[\prod_{d=1}^{D} \theta_{dk}^{x_{nd}} (1-\theta_{dk})^{1-x_{nd}} \right]^{r_{nk}}$

2. Write the expression for the joint distribution $p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N|\boldsymbol{\theta}, \mathbf{r})$.

A. $p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N|\boldsymbol{\theta}, \mathbf{r}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\prod_{d=1}^{D} \theta_{dk}^{x_{nd}} (1-\theta_{dk})^{1-x_{nd}} \right]^{r_{nk}}$

3. Derive the marginal distribution $p(\mathbf{x}_n|\boldsymbol{\theta}, \boldsymbol{\pi})$.

A. $p(\mathbf{x}_n|\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{k=1}^{K} p(\mathbf{x}_n, r_n = k|\boldsymbol{\theta}) = \sum_{k=1}^{K} p(\mathbf{x}_n|r_n = k, \boldsymbol{\theta})\pi_k = \sum_{k=1}^{K} \prod_{d=1}^{D} \theta_{dk}^{x_{nd}} (1-\theta_{dk})^{1-x_{nd}} \pi_k$

4. Derive the posterior distribution $p(r_n = k|\mathbf{x}_n, \boldsymbol{\theta}, \boldsymbol{\pi})$.

A. We use the Bayes rule to write the following:

$$p(r_n = k|\mathbf{x}_n, \boldsymbol{\theta}, \boldsymbol{\pi}) = \frac{p(\mathbf{x}_n|r_n = k, \boldsymbol{\theta}, \boldsymbol{\pi}) p(r_n = k)}{p(\mathbf{x}_n|\boldsymbol{\theta}, \boldsymbol{\pi})} = \frac{\prod_{d=1}^{D} \theta_{dk}^{x_{nd}} (1-\theta_{dk})^{1-x_{nd}} \pi_k}{\sum_{j=1}^{K} \prod_{d=1}^{D} \theta_{dj}^{x_{nd}} (1-\theta_{dj})^{1-x_{nd}} \pi_j} \tag{25}$$

5. Write the expression for maximum likelihood estimator.

A. Using the expression for the marginal $p(\mathbf{x}_n|\boldsymbol{\theta}, \boldsymbol{\pi})$, we get

$$\max_{\boldsymbol{\theta}, \boldsymbol{\pi}} \log \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\theta}, \boldsymbol{\pi}) = \max_{\boldsymbol{\theta}, \boldsymbol{\pi}} \sum_{n=1}^{N} \log p(\mathbf{x}_n|\boldsymbol{\theta}, \boldsymbol{\pi}) = \max_{\boldsymbol{\theta}, \boldsymbol{\pi}} \sum_{n=1}^{N} \log \sum_{k=1}^{K} \prod_{d=1}^{D} \theta_{dk}^{x_{nd}} (1-\theta_{dk})^{1-x_{nd}} \pi_k \tag{26}$$

6. Do you think that the cost function is jointly-convex? identifiable?

A. Following the same argument as GMM (and K-means), we can say that the function is not jointly convex and also that it is not identifiable.

7. Derive a lower bound to $\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\theta})$ in the E-step using Jensen's inequality.

A. Simply following the GMM derivation,

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\theta}) &:= \sum_{n=1}^{N} \log \sum_{k=1}^{K} \prod_{d=1}^{D} \theta_{dk}^{x_{nd}} (1-\theta_{dk})^{1-x_{nd}} \pi_k && (27)\\
&= \sum_{n=1}^{N} \log \left[\sum_{k=1}^{K} \frac{\prod_{d=1}^{D} \theta_{dk}^{x_{nd}} (1-\theta_{dk})^{1-x_{nd}} \pi_k}{p_{kn}^{(i)}} p_{kn}^{(i)} \right] && (28)\\
&\geq \sum_{n=1}^{N} \sum_{k=1}^{K} \log \left[\frac{\prod_{d=1}^{D} \theta_{dk}^{x_{nd}} (1-\theta_{dk})^{1-x_{nd}} \pi_k}{p_{kn}^{(i)}} \right] p_{kn}^{(i)} && (29)\\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{d=1}^{D} \log \left[\theta_{dk}^{x_{nd}} (1-\theta_{dk})^{1-x_{nd}} \pi_k \right] p_{kn}^{(i)} - \text{cnst} && (30)\\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{d=1}^{D} \left[x_{nd} \log \theta_{dk} + (1-x_{nd}) \log(1-\theta_{dk}) + \log \pi_k \right] p_{kn}^{(i)} - \text{cnst} && (31)
\end{aligned}$$

Note: We could skip the first few steps and use Eq. (19) in Page 17 of GMM lecture notes.

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\theta}) &:= \sum_{n} \log p(\mathbf{x}_n|\boldsymbol{\pi}, \boldsymbol{\theta}) \geq \sum_{n} \mathbb{E}_{p_{kn}^{(i)}} \left[\log p(\mathbf{x}_n, r_n|\boldsymbol{\pi}, \boldsymbol{\theta}) \right] + \text{cnst} && (32)\\
&= \sum_{n} \sum_{k=1}^{K} \log \left[\frac{\prod_{d=1}^{D} \theta_{dk}^{x_{nd}} (1-\theta_{dk})^{1-x_{nd}} \pi_k}{p_{kn}^{(i)}} \right] p_{kn}^{(i)} && (33)
\end{aligned}$$

8. Derive the M-step update for $\theta_{dk}, \forall d, k$ by maximizing the lower bound obtained in the E-step.

A. Taking derivative of the lower bound wrt θ_{dk}, we get,

$$\sum_{n=1}^{N} \left[\frac{x_{nd}}{\theta_{dk}} - \frac{1 - x_{nd}}{1 - \theta_{dk}} \right] p_{kn}^{(i)} = 0 \tag{34}$$

$$\Rightarrow \sum_{n=1}^{N} \left[x_{nd}(1 - \theta_{dk}) - (1 - x_{nd})\theta_{dk} \right] p_{kn}^{(i)} = 0 \tag{35}$$

$$\Rightarrow \sum_{n=1}^{N} \left[x_{nd} - \theta_{dk} \right] p_{kn}^{(i)} = 0 \tag{36}$$

$$\Rightarrow \theta_{dk} = \frac{\sum_{n=1}^{N} x_{nd} p_{kn}^{(i)}}{\sum_{n=1}^{N} p_{kn}^{(i)}} \tag{37}$$

Notice the similarity to GMM updates!

9. Do you think that the EM updates will return ....

A. We can clearly see that each update will be within the range, therefore yes the update will reutrn a valid value.

Bayesian linear regression Derive expressions for the posterior distribution $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$ and the marginal likelihood $p(\mathbf{y}|\mathbf{X})$.

A. The prior and the likelihood can be read from the joint.

$$p(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}|0, \mathbf{I}) \tag{38}$$

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \mathbf{I}) \tag{39}$$

Using the formula, we get the following distributions:

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|0, \mathbf{I} + \mathbf{X}\mathbf{X}^T) \tag{40}$$

$$p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\Sigma}\mathbf{X}^T\mathbf{y}, \boldsymbol{\Sigma}) \tag{41}$$

$$\boldsymbol{\Sigma} = (\mathbf{I} + \mathbf{X}^T\mathbf{X})^{-1} \tag{42}$$

Note: This calculation might seem straightforward but these formula will be useful in the future.

PCA for count data 1. Derive the log-likelihood for this model.

A. $\log p(\mathbf{X}|\mathbf{W}, \mathbf{Z}) = \sum_n \sum_d \left[x_{nd} \mathbf{w}_d^T \mathbf{z}_n - \exp(\mathbf{w}_d^T \mathbf{z}_n) \right] + \text{cnst}$

2. Show that $\mathcal{L}$ is convex with respect with respect to $\mathbf{W}$ given $\mathbf{Z}$ and vice-versa.

A. Given all $\mathbf{z}_n$, the cost function wrt $\mathbf{w}_d$ is the following: $-\sum_n x_{nd} \mathbf{w}_d^T \mathbf{z}_n + \exp(\mathbf{w}_d^T \mathbf{z}_n)$. Each term in this is convex wrt $\mathbf{w}_d$ since the first term is linear and second term is convex since $\exp(\cdot)$ is convex. Since sum of convex functions is convex, the resulting function is also convex.

3. Is this model identifiable? Why and why not? Discuss your answer.

A. Similar to PCA, we can multiply the two factors by constants a and $1/a$ respectively and get the same likelihood value. Therefore the model is not identifiable.

4. Write an algorithm similar to alternating least-squares ...

A. Given all $\mathbf{z}_n$, the cost function wrt $\mathbf{w}_d$ is the following: $-\sum_n x_{nd} \mathbf{w}_d^T \mathbf{z}_n + \exp(\mathbf{w}_d^T \mathbf{z}_n)$. Minimizing this function is equivalent to Poisson regression for which there is no closed form solution. We must therefore run gradient descent. Here is an algorithm based on this.

(a) Intialize $\mathbf{W}$ and $\mathbf{Z}$.

(b) For each row w_d of $\mathbf{W}$, by running the following steps. Stop when cost function changes less than some $\epsilon > 0$.

i. $\mathbf{w}_d \leftarrow \mathbf{w}_d - \alpha \sum_n \mathbf{z}_n \left[-x_{nd} + \exp(\mathbf{w}_d^T \mathbf{z}_n)\right]$.

(c) For each row $\mathbf{z}_n$ of $\mathbf{Z}$, by running the following steps. Stop when cost function changes less than some $\epsilon > 0$.

i. $\mathbf{z}_n \leftarrow \mathbf{z}_n - \alpha \sum_d \mathbf{w}_d \left[-x_{nd} + \exp(\mathbf{w}_d^T \mathbf{z}_n)\right]$.

(d) Check convergence using the negative of the log-likelihood.

5. What is the computational complexity of the algorithm?

A. $O(M^2DN)$, ignoring the number of iterations in each sub-step when optimizing $\mathbf{z}_n$ and $\mathbf{w}_d$.

6. What would you do to reduce overfitting? Why?

A. We can add penalty terms $\mathbf{w}_d^T \mathbf{w}_d$ and $\mathbf{z}_n^T \mathbf{z}_n$. Each regression step will reduce overfitting similar to ridge regression or penalized logistic regression.

7. Is it possible to obtain a closed-form solution using SVD? Why and why not? Discuss your answer.

A. It is not possible to get a closed-form expression since the factorization is passed through a non-linear function.

Naive Bayes classifier Answer the following questions.

1. Draw the graph which corresponds to this factorization.

A. The graph is $x_1 \leftarrow y \rightarrow x_2$.

2. Compute the following posterior values.

(a) $p(y = 1|x_1 = 1, x_2 = 1)$

(b) $p(y = 1|x_1 = 1, x_2 = 0)$

(c) $p(y = 1|x_1 = 0, x_2 = 1)$

(d) $p(y = 1|x_1 = 0, x_2 = 0)$

A. I will demonstrate the answer for the first part only, i.e. to compute $p(y = 1|x_1 = 1, x_2 = 1)$. Using Bayes rule,

$$p(y = 1|x_1 = 1, x_2 = 1) \propto p(x_1 = 1|y = 1)p(x_2 = 1|y = 1)p(y = 1) = 0.2 * 0.5 * 0.5 \quad (43)$$

$$p(y = 0|x_1 = 1, x_2 = 1) \propto p(x_1 = 1|y = 0)p(x_2 = 1|y = 0)p(y = 0) = 0.9 * 0.5 * 0.5 \quad (44)$$

Therefore,

$$p(y = 1|x_1 = 1, x_2 = 1) = 0.2/1.1 \approx 0.18 \quad (45)$$

Similarly, you can compute other values.

$$p(y = 1|x_1 = 1, x_2 = 0) \approx 0.18 \quad (46)$$

$$p(y = 1|x_1 = 0, x_2 = 1) \approx 0.89 \quad (47)$$

$$p(y = 1|x_1 = 0, x_2 = 0) \approx 0.89 \quad (48)$$

3. Let us say that you have to make your decision (whether $y = 1$ or not) based on either x_1 or x_2, i.e. you have to choose one of those. Which one will you choose?

A. It is easy to see in the posterior that x_2 does not affect the posterior, so we can discard it.

Kernels Show that the following function is a Kernel and derive the basis function $\boldsymbol{\phi}(\mathbf{x})$ ...

A. To prove that the function is Kernel, we need two properties: symmetry and positive-semi-definiteness (p.s.d.). We can easily see that the function is symmetric.

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j)^2 = (\mathbf{x}_j^T \mathbf{x}_i)^2 = K(\mathbf{x}_j, \mathbf{x}_i) \tag{49}$$

To prove p.s.d. property, we need that the product $\mathbf{t}^T\mathbf{K}\mathbf{t} \geq 0$ for all non-zero vectors $\mathbf{t}$ ($\mathbf{K}$ is the $N \times N$ matrix formed with N features). To prove this, we rewrite the product as follows,

$$\mathbf{t}^T\mathbf{K}\mathbf{t} = \sum_i \sum_j K_{ij} t_i t_j = \sum_i \sum_j (\mathbf{x}_i^T \mathbf{x}_j)^2 t_i t_j = \sum_i \sum_j (\mathbf{x}_i^T \mathbf{x}_j)(\mathbf{x}_i^T \mathbf{x}_j) t_i t_j$$

$$= \sum_i \sum_j \left(\sum_k x_{ik} x_{jk} \right) \left(\sum_l x_{il} x_{jl} \right) t_i t_j = \sum_k \sum_l \left(\sum_i t_i x_{il} x_{ik} \right) \left(\sum_j t_j x_{jl} x_{jk} \right) \tag{50}$$

$$= \sum_k \sum_l \left(\sum_i t_i x_{il} x_{ik} \right)^2 > 0 \tag{51}$$

Hence proved.

Note: You will get marks for an attempt even if you are unable to prove.

To find a feature map, we will proceed recursively. Let's rename $\mathbf{x}_i$ and $\mathbf{x}_j$ as $\mathbf{x}$ and $\mathbf{y}$ for notational convenience. We start with 2-D vector,

$$(x_1 y_1 + x_2 y_2)^2 = x_1^2 y_1^2 + x_1^2 y_1^2 + 2 x_1 x_2 y_1 y_2 = \begin{bmatrix} x_1^2 & x_2^2 & \sqrt{2} x_1 x_2 \end{bmatrix} \begin{bmatrix} y_1^2 \\ y_2^2 \\ \sqrt{2} y_1 y_2 \end{bmatrix} \tag{52}$$

Then 3-D vector,

$$(x_1 y_1 + x_2 y_2 + x_3 y_3)^2 = (x_1 y_1 + x_2 y_2)^2 + x_3^2 y_3^2 + 2(x_1 y_1 + x_2 y_2) x_3 y_3 \tag{53}$$

$$= \begin{bmatrix} x_1^2 & x_2^2 & \sqrt{2} x_1 x_2 & x_3^2 & \sqrt{2} x_1 x_3 & \sqrt{2} x_2 x_3 \end{bmatrix} \begin{bmatrix} y_1^2 \\ y_2^2 \\ \sqrt{2} y_1 y_2 \\ y_3^2 \\ \sqrt{2} y_3 \\ \sqrt{2} y_2 y_3 \end{bmatrix} \tag{54}$$

If we proceed in this way, we get the following features:

$$\boldsymbol{\phi}(\mathbf{x}) = [x_1^2, x_2^2, \ldots, x_D^2, \sqrt{2} x_1 x_2, \sqrt{2} x_1 x_3, \ldots, \sqrt{2} x_1 x_D, \sqrt{2} x_2 x_3, \sqrt{2} x_2 x_4, \ldots, \sqrt{2} x_2 x_D, \ldots\ldots]^T \tag{55}$$

Artificial neural networks Answer the following questions.

1. Write down the forward equations to compute the activations, hidden units and the output.

A. For the first layer,

$$a_{n1}^{(1)} = \widetilde{\mathbf{x}}_n^T \boldsymbol{\beta}_1^{(1)} \quad , \quad z_{n1}^{(1)} = \sigma(a_{n1}^{(1)}) \tag{56}$$

$$a_{n2}^{(1)} = \widetilde{\mathbf{x}}_n^T \boldsymbol{\beta}_2^{(1)} \quad , \quad z_{n2}^{(1)} = \sigma(a_{n2}^{(1)}) \tag{57}$$

$$a_{n3}^{(1)} = \widetilde{\mathbf{x}}_n^T \boldsymbol{\beta}_3^{(1)} \quad , \quad z_{n3}^{(1)} = \sigma(a_{n3}^{(1)}) \tag{58}$$

$$a_{n4}^{(1)} = \widetilde{\mathbf{x}}_n^T \boldsymbol{\beta}_4^{(1)} \quad , \quad z_{n4}^{(1)} = \sigma(a_{n4}^{(1)}) \tag{59}$$

For the second layer (assuming a real value output),

$$a_{n1}^{(2)} = \widetilde{\mathbf{z}}_n^T \boldsymbol{\beta}_1^{(2)} \quad , \quad z_{n1}^{(2)} = \sigma(a_{n1}^{(2)} \quad , \quad \hat{y}_n = z_{n1}^{(2)} \tag{60}$$

2. What is the total number of parameters in this model?

A. Each activation in the first layer involves a parameter vector $\boldsymbol{\beta}_m$ of length 4, so in total there are $4 \times 4 = 16$ parameters in the first layer. The second layer has 5 parameters, so in total we have 21 parameters.

3. Write down the gradient of $\mathcal{L}_n$ with respect to the parameter $\beta_{34}^{(1)}$ ...

A. We will use the back-propagation rule. There is a small typo in the lecture notes, so I wrote the equations here again. Similar to the lecture, we ignore subscript n. Let $\boldsymbol{\delta}^{(k)} = \frac{\partial \mathcal{L}}{\partial \mathbf{a}^{(k)}}$, then the following backpropagation algorithm can be used to compute the derivatives:

$$\boldsymbol{\delta}^{(k-1)} = \text{diag}[\mathbf{h}'(\mathbf{a}^{(k-1)})]\mathbf{B}^{(k)T}\boldsymbol{\delta}^{(k)} \tag{61}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{B}^{(1)}} = \boldsymbol{\delta}^{(1)}\mathbf{x}^T \tag{62}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{B}^{(k)}} = \boldsymbol{\delta}^{(k)}\mathbf{z}^{(k)T} \tag{63}$$

where $\mathbf{B}^{(k)}$ is $M \times (M+1)$ matrix containing $\boldsymbol{\beta}_m^{(k)T}$ as rows.

Another important point to note here is that I am denoting the elements of $\boldsymbol{\beta}_m^{(k)}$ by $\beta_{dm}^{(k)}$ (d is the first subscript).

We require to compute derivative wrt $\beta_{34}^{(1)}$ of the first layer. We can compute this derivative using Eq. 62. The matrix $\partial\mathcal{L}/\partial\mathbf{B}^{(1)}$ is of size 4×4 and the entry that correspond to the derivative of $\beta_{34}^{(1)}$ is in the last row and last column. This entry corresponds to $\delta_4^{(1)}x_3$. Therefore, we need to compute $\delta_4^{(1)}$.

Note: Make sure to write the full matrix form to understand how we arrive at this equation.

We start backpropagation by computing $\frac{\partial \mathcal{L}}{\partial a_1^{(2)}}$ (there is only one activation in the second layer).

$$\frac{\partial \mathcal{L}}{\partial a_1^{(2)}} = \frac{\partial}{\partial a_1^{(2)}}\tfrac{1}{2}[y - \sigma(a_1^{(2)})]^2 = [\sigma(a_1^{(2)}) - y]\frac{\partial \sigma(a_1^{(2)})}{\partial a_1^{(2)}} \tag{64}$$

$$= [\sigma(a_1^{(2)}) - y]\sigma(a_1^{(2)})[1 - \sigma(a_1^{(2)})] \tag{65}$$

Here, we have added a $\frac{1}{2}$ in MSE for notational convenience. The last step is obtained using the fact that derivative of $\sigma(a)$ is $\sigma(a)[1 - \sigma(a)]$.

Next, we use the recursion of Eq. 61 to compute the following $\partial\mathcal{L}/\partial a_4^{(1)}$. This is basically the last element of $\boldsymbol{\delta}^{(1)}$ which can be computed as following:

$$\frac{\partial \mathcal{L}}{\partial a_4^{(1)}} = h'(a_4^{(1)})\beta_{41}^{(2)}\frac{\partial \mathcal{L}}{\partial a_1^{(2)}} \tag{66}$$

$$= \sigma(a_4^{(1)})[1 - \sigma(a_4^{(1)})]\beta_{41}^{(2)}\frac{\partial \mathcal{L}}{\partial a_1^{(2)}} \tag{67}$$

Note: Make sure to write the full matrix form to understand how we arrive at this equation.

Using the above equation, we can now write the gradient.

$$\frac{\partial \mathcal{L}}{\partial \beta_{34}^{(1)}} = x_3\delta_4^{(1)} = x_3\sigma(a_4^{(1)})[1 - \sigma(a_4^{(1)})]\beta_{41}^{(2)}\frac{\partial \mathcal{L}}{\partial a_1^{(2)}} \tag{68}$$

$$= x_3\sigma(a_4^{(1)})[1 - \sigma(a_4^{(1)})]\beta_{41}^{(2)}[\sigma(a_1^{(2)}) - y]\sigma(a_1^{(2)})[1 - \sigma(a_1^{(2)})] \tag{69}$$

Bayesian networks and Belief propagation Answer the following questions.

1. Write the factorization of the joint $p(y_1, y_2, z_1, z_2, z_3)$ under the Bayesian network.

A. $p(y_1|z_1, z_2, z_3)p(y_2|z_3)p(z_1)p(z_2)p(z_3)$.

2. Write down the expression to compute the marginals from the factorized-joint distribution.

A. $p(z_1)\sum_{z_2} p(z_2)\sum_{z_3} p(y_1|z_1, z_2, z_3)p(y_2|z_3)p(z_3)$.

3. Write down the expressions for the messages from the variables to the observations.

A. To avoid confusiion, we will call y_1 as y_a and y_2 as y_b.

$$m_{1\to a}(z_1) = p(z_1) \tag{70}$$
$$m_{2\to a}(z_2) = p(z_2) \tag{71}$$
$$m_{3\to a}(z_3) = p(z_3)m_{b\to 3}(z_3) \tag{72}$$
$$m_{3\to b}(z_3) = p(z_3)m_{a\to 3}(z_3) \tag{73}$$

4. Write down the expressions for the messages from the observations to the variables.

A.

$$m_{a\to 1}(z_1) = \sum_{z_2}\sum_{z_3} p(y_a|z_1, z_2, z_3)m_{2\to a}(z_2)m_{3\to a}(z_3) \tag{74}$$
$$m_{a\to 2}(z_2) = \sum_{z_1}\sum_{z_3} p(y_a|z_1, z_2, z_3)m_{1\to a}(z_1)m_{3\to a}(z_3) \tag{75}$$
$$m_{a\to 3}(z_3) = \sum_{z_1}\sum_{z_2} p(y_a|z_1, z_2, z_3)m_{1\to a}(z_1)m_{2\to a}(z_2) \tag{76}$$
$$m_{b\to 3}(z_3) = p(y_b|z_3) \tag{77}$$

5. Which path computes the marginal $p(z_1|y_1, y_2)$?

A. The path is the following: $b \to 3 \to a$ then $2 \to a$ then $a \to 1$.

$$p(z_1|y_a, y_b) \propto p(z_1) \underbrace{\sum_{z_2}\sum_{z_3} p(y_a|z_1, z_2, z_3) \underbrace{p(z_2)}_{m_{2\to a}(z_2)} \underbrace{p(z_3)\underbrace{p(y_b|z_3)}_{m_{b\to 3}(z_3)}}_{m_{3\to a}(z_3)}}_{m_{a\to 1}(z_1)} \tag{78}$$

Note: Finding the path could be a tricky. You should work your path on the graph itself because that makes it easier. Start the message flow from the end points and work towards the end point. Once you write a path, try to work the messages in the marginal expression. If they don't match, you might have a problem in the path (or may be the expression itself).