PCML 2014: Sample Exam Questions

Mohammad Emtiyaz Khan
EPFL

January 4, 2015

Abstract

I have added a few questions to give you an idea of the type of questions the exam will contain. I will be updating these questions and the updated date is shown along with every question. I have also indicated difficulty level on three levels: easy, moderate, difficult. The solutions to these questions will be available around 7 days before the final exam.

1 Information about the exam

Few important information about the exam.

- You will have 180 minutes in total.
- We will provide you extra work sheets, but you will have to submit them along with your exam.
- You are not allowed to use books or any other help material except an A4 size cheat-sheet (can use both sides). Note that you also have to submit your cheat sheet along with your exam.
- No electronic devices are allowed except a calculator. Make sure that your calculator is only a calculator and cannot be used for any other purpose than for manual numerical-calculations.
- := means "defined as".
- For derivations, clearly explain your derivation step by step. You will be marked for steps as well as for the end result.
- The exam has a total of 60 points

2 Questions

Weighted least-squares : [Updated on Dec. 14, 2014] [Easy]

Suppose we have a regression dataset with N pairs $\{y_n, \mathbf{x}_n\}$ where y_n is a real-valued scalar, $\mathbf{x}_n$ is a real-valued vector of length D, and we wish to fit a linear model $f(\mathbf{x}_n) = \boldsymbol{\beta}^T \widetilde{\mathbf{x}}_n$ where $\boldsymbol{\beta}$ is a vector with entries $\beta_0, \beta_1, \ldots, \beta_D$ and $\widetilde{\mathbf{x}}_n^T = [1\, \mathbf{x}_n^T]$.

Suppose we minimize the following cost function:

$$\mathcal{L}(\boldsymbol{\beta}) = \tfrac{1}{2} \sum_{n=1}^{N} w_n \left(y_n - \boldsymbol{\beta}^T \widetilde{\mathbf{x}}_n \right)^2 \tag{1}$$

where $w_n > 0$ are known real-valued scalars.

1

$$\mathcal{L}(\beta) = \frac{1}{2}\sum w_n (y_n - \beta^T \tilde{x}_n)^2$$

$$\frac{\partial \mathcal{L}(\beta)}{\partial \beta} = -\sum \tilde{x}_n^T w_n (y_n - \beta^T \tilde{x}_n)$$

$$- X^T W^T (y_n + \beta^T \tilde{X}_n) = 0$$

$$- X^T W y_n + X^T W \beta^T X_n = 0$$

2

$$X^T W X_n \beta = X^T W y_n$$

$$\beta^* = (X^T W X)^{-1} X^T y_n$$

$X^T W X$ – invertible

$W > 0$

so X must be full rank.

(a) Derive the normal equations for this cost function. You should write an expression in matrix-vector form similar to the expression for least-squares given in the lecture notes.

(b) Discuss the conditions under which the solution $\boldsymbol{\beta}^*$ is unique. as the function is convex an w is an scaler, the $f(p)$ is convex so is unique

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

(d) We showed in the lectures that the least-squares cost function can be derived using a probabilistic model. Derive a probabilistic model under which minimizing the negative of the log-likelihood gives the same solution as the cost function shown above in Eq. 1. $p(y|X,\beta,w) = X\beta + \varepsilon$

$\varepsilon \sim N(0|X\beta, w)$

Multi-class classification: [Updated on Dec. 15, 2014] [Moderate]

[Update on Dec. 28, 2014] Changed k to j in Eq. 2.

Suppose we have a classification dataset with N pairs $\{y_n, \mathbf{x}_n\}$ but now y_n is a categorical variable, i.e. $y_n \in \{1, 2, \ldots, K\}$ where K is the number of classes. We wish to fit a linear model and in the similar spirit to logistic regression, we will use a multinomial logit distribution to map linear inputs to a categorical output.

We will define $\eta_{nk} = \widetilde{\mathbf{x}}_n^T \boldsymbol{\beta}_k$ for all $k = 1, 2, \ldots, K-1$ and then compute the probability of output,

$$p(y_n = k|\mathbf{x}_n, \boldsymbol{\beta}) = \frac{e^{\eta_{nk}}}{\sum_{j=1}^{K} e^{\eta_{nj}}} \tag{2}$$

For identifiability reasons, we set $\eta_{nK} = 0$, therefore $\boldsymbol{\beta}_K = 0$ and we need to estimate $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots = \beta_{K-1}$.

Similar to logistic regression, we will assume that each y_n is i.i.d. i.e.

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \prod_{n=1}^{N} p(y_n = k|\mathbf{x}_n, \boldsymbol{\beta}) \tag{3}$$

Following the derivation of logistic regression,

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

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

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

Poisson regression : [Updated on Dec. 15, 2014] [Moderate]

Suppose we have a regression dataset with N pairs $\{y_n, \mathbf{x}_n\}$ but now y_n is an integer valued scalar, i.e. $y_n \in \{1, 2, 3, 4, \ldots\}$. We wish to fit a linear model $f(\mathbf{x}_n) = \boldsymbol{\beta}^T \widetilde{\mathbf{x}}_n$.

In the similar spirit to logistic regression, we will use a Poisson distribution to map $\eta_n = \widetilde{\mathbf{x}}_n^T \boldsymbol{\beta}$ to an integer output, i.e. we define the following probability,

$$p(y_n = k|\mathbf{x}_n, \boldsymbol{\beta}) = \frac{e^{k\eta_n}}{k!} e^{-e^{\eta_n}} \tag{4}$$

Also, similar to logistic regression, we will assume that each y_n is i.i.d. i.e.

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \prod_{n=1}^{N} p(y_n = k|\mathbf{x}_n, \boldsymbol{\beta}) \tag{5}$$

Following the derivation of logistic regression,

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

(b) Derive the normal equations.

(c) Derive the Hessian.

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

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

EM for mixture of Bernoulli [Updated on Dec. 26, 2014] [Difficult]

[Update on Dec. 31, 2014] Small rewording of part 7.

Consider the problem of clustering N vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ where each $\mathbf{x}_n$ is **binary** vector of length D. To model the binary data, we can use a mixture of Bernoulli distribution with K mixtures. Similar to GMM, we can assume that each vector $\mathbf{x}_n$ is independently distributed but now the k'th mixture component models the data vector $\mathbf{x}_n$ using a Bernoulli distribution, as shown below.

$$p(\mathbf{x}_n | r_n = k, \boldsymbol{\theta}) = \prod_{d=1}^{D} p(x_{nd}|\theta_{dk}) := \prod_{d=1}^{D} \theta_{dk}^{x_{nd}} (1-\theta_{dk})^{1-x_{nd}} \tag{6}$$

where r_n indicate the cluster assignment, $\theta_{dk} \in (0,1)$ is the probability that x_{dn} takes a value 1 for cluster k (exactly like logistic regression). Let $\boldsymbol{\theta}_k = [\theta_{1k}, \theta_{2k}, \dots, \theta_{Dk}]^T$ be the parameter vector for cluster k (similar to $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ in GMM) and $\boldsymbol{\theta}$ be the vector containing all $\boldsymbol{\theta}_k$.

Answer the following questions. For derivations, clearly show each step since you will be marked for steps, not only for the end result.

1. Similar to lecture notes, define $\mathbf{r}_n$ to be a binary vector of length K where an entry of 1 at k'th position indicates that the n'th data vector belong to k'th mixture and implies that $r_n = k$. Rewrite the likelihood $p(\mathbf{x}_n|\boldsymbol{\theta}, \mathbf{r})$ in terms of r_{nk} (your answer should be similar to Eq 2 in GMM lecture notes).
2. Let $\mathbf{r}$ be a vector containing all $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$. Write the expression for the joint distribution $p(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N|\boldsymbol{\theta}, \mathbf{r})$.
3. Similar to GMM, let us assume that r_n follows a multinomial distribution $p(r_n = k|\boldsymbol{\pi}) = \pi_k$. Derive the marginal distribution $p(\mathbf{x}_n|\boldsymbol{\theta}, \boldsymbol{\pi})$ (similar to Eq. 4 in GMM lecture notes) where $\boldsymbol{\pi}$ is the vector containing all π_k.
4. Derive the posterior distribution $p(r_n = k|\mathbf{x}_n, \boldsymbol{\theta}, \boldsymbol{\pi})$ (similar to Eq. 5 in GMM lecture notes).
5. Write the expression for maximum likelihood estimator in terms of data vectors $\mathbf{x}_n$ and parameters $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ (similar to Eq. 6 in GMM lecture notes).
6. Define the cost function to be the negative of maximum likelihood. Do you think that the cost function is jointly-convex with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$? Is the model identifiable?
7. Consider the maximum likelihood cost function $\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\theta})$. Let us say that we are at the i'th iteration of an EM algorithm with current parameter estimates $\boldsymbol{\theta}^{(i)}$ and $\boldsymbol{\pi}^{(i)}$. Denote the posterior $p_{kn}^{(i)} := p(r_n = k|\mathbf{x}_n, \boldsymbol{\theta}^{(i)}, \boldsymbol{\pi}^{(i)})$. Derive a lower bound to $\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\theta})$ in the E-step using Jensen's inequality. The lower bound should be a function of $p_{kn}^{(i)}, \boldsymbol{\pi}, \boldsymbol{\theta}$.
8. Derive the M-step update for $\theta_{dk}, \forall d, k$ by maximizing the lower bound obtained in the E-step.
9. Each θ_{dk} should be in the range $(0,1)$. Do you think that the EM updates will return a value that is (strictly) greater than 1 or (strictly) less than 0? Why and why not?

Bayesian linear regression [Updated on Dec. 26, 2014] [Easy]

Consider the following joint model for linear regression:

$$p(\mathbf{y}, \boldsymbol{\beta}|\mathbf{X}) = \left[\prod_{n=1}^{N} \mathcal{N}(y_n|\boldsymbol{\beta}^T\mathbf{x}_n, 1)\right] \mathcal{N}(\boldsymbol{\beta}|0, \mathbf{I}) \tag{7}$$

Here, we do not consider the bias parameter β_0 and $\boldsymbol{\beta}$ is therefore of the same length as $\mathbf{x}_n$.

The Gaussian formula from Bishop Chapter 2 (Eq. 2.113 to 2.117) compute the posterior and marginal likelihood. Using these formula, derive expressions for the posterior distribution $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$ and the marginal likelihood $p(\mathbf{y}|\mathbf{X})$.

PCA for count data [Updated on Dec. 26, 2014] [Easy]

Given $D \times N$ data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, we can compute a low-rank approximation such that $\mathbf{X} \approx \mathbf{W}\mathbf{Z}^T$. In the class, we estimated $\mathbf{W}$ and $\mathbf{Z}$ by minimizing the reconstruction MSE using an alternative least-squares algorithm. This is equivalent to computing a maximum likelihood estimate for a model with the following likelihood:

$$p(\mathbf{X}|\mathbf{W}, \mathbf{Z}) = \prod_{n=1}^{N} \prod_{d=1}^{D} \mathcal{N}(x_{nd}|\mathbf{w}_d^T \mathbf{z}_n, 1) \tag{8}$$

Here, $\mathbf{w}_d$ is d'th row of $\mathbf{W}$ and $\mathbf{z}_n$ is n'th row of $\mathbf{Z}$. Make sure to prove that maximum likelihood estimation with this data likelihood is equivalent to minimizing reconstruction MSE.

Let us now assume that, similar to music recommendation data for project 2, $\mathbf{X}$ contains only integer values, i.e. each $x_{nd} \in \{0, 1, 2, \ldots\}$. Unlike your project 2 data, an entry of 0 does not mean a missing value in this case rather means a count of 0. For such count data, we can assume a Poisson distribution likelihood as shown below.

$$p(\mathbf{X}|\mathbf{W}, \mathbf{Z}) = \prod_{n=1}^{N} \prod_{d=1}^{D} \frac{e^{x_{nd}\eta_{nd}}}{x_{nd}!} e^{-e^{\eta_{nd}}}, \qquad \text{where } \eta_{nd} = \mathbf{w}_d^T \mathbf{z}_n \tag{9}$$

1. Derive the log-likelihood for this model.
2. Define the cost function $\mathcal{L}(\mathbf{W}, \mathbf{Z})$ to be the negative of log-likelihood. Show that $\mathcal{L}$ is convex with respect with respect to $\mathbf{W}$ given $\mathbf{Z}$ and vice-versa.
3. Is this model identifiable? Why and why not? Discuss your answer.
4. Write an algorithm similar to alternating least-squares where we iterate between optimizing over $\mathbf{W}$ given $\mathbf{Z}$ and then optimizing over $\mathbf{Z}$ given $\mathbf{W}$. Discuss the convergence criteria of each iteration as well as the whole algorithm?
5. What is the computational complexity of the algorithm?
6. What would you do to reduce overfitting? Why?
7. Is it possible to obtain a closed-form solution using SVD? Why and why not? Discuss your answer.

Naive Bayes classifier [Updated on Dec. 28, 2014] [Easy]

[Update on Jan. 4, 2015] Corrected a mistake in Eq. 13.

Consider a binary classification problem with one binary output y and two binary features x_1 and x_2. Naive Bayes classifier assumes the following joint distribution for a pair,

$$p(y, x_1, x_2) = p(x_1|y)p(x_2|y)p(y) \tag{10}$$

Let the values of these probabilities be:

$$p(y = 0) = 0.5 \qquad p(y = 1) = 0.5 \tag{11}$$

$$p(x_1 = 1|y = 0) = 0.9 \qquad p(x_1 = 1|y = 1) = 0.2 \tag{12}$$

$$p(x_2 = 1|y = 0) = 0.5 \qquad p(x_2 = 1|y = 1) = 0.5 \tag{13}$$

Answer the following questions.

1. Draw the graph which corresponds to this factorization.
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) $p(y = 1|x_1 = 0, x_2 = 0)$
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?

Kernels [Updated on Dec. 28, 2014] [Easy]

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

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j)^2 \tag{14}$$

Artificial neural networks [Updated on Dec. 29, 2014] [Easy]

Consider the following artificial neural network with the nonlinear transformation $z_{nm} = \sigma(a_{nm})$ (see Fig. 1). Here, n is the data index and m is the index of hidden units (as described in the lecture notes). We will denote the parameter with the symbol $\boldsymbol{\beta}$.

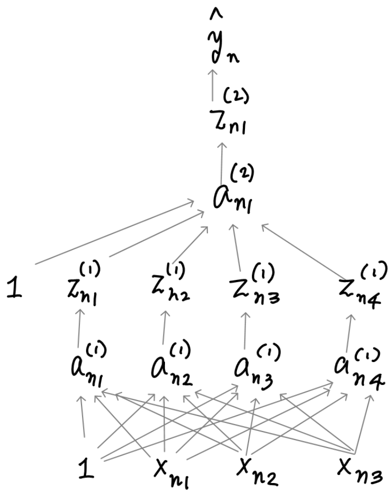

Figure 1: Artificial neural network

Answer the following questions.

1. Write down the forward equations to compute the activations, hidden units and the output.
2. What is the total number of parameters in this model? Show your answer clearly counting the number of parameters for each activation.
3. Suppose we use a MSE cost function for the n'th data example: $\mathcal{L}_n(\boldsymbol{\beta}) := (y_n - \hat{y}_n)^2$. Write down the gradient of $\mathcal{L}_n$ with respect to the parameter $\beta^{(1)}_{34}$ that connects the input x_{n3} to the activation $a^{(1)}_{n4}$. You should use the chain rule as we learned in the backpropagation algorithm.

Bayesian networks and Belief propagation [Updated on Dec. 29, 2014] [Difficult]

Suppose that we have the Bayesian network shown in Fig. 2.

Answer the following questions.

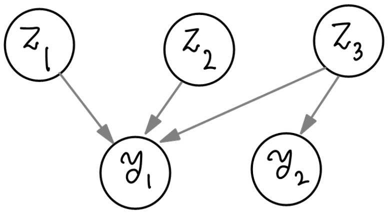

Figure 2: A Bayesian network

1. Write the factorization of the joint $p(y_1, y_2, z_1, z_2, z_3)$ under the Bayesian network.
2. Suppose we want to compute the marginals $p(z_1|y_1, y_2), p(z_2|y_1, y_2)$ and $p(z_3|y_1, y_2)$. Write down the expression to compute the marginals from the factorized-joint distribution. Clearly show the factorization of the joint. You must also push the sums inside as described in sum-product lecture. Your expression should resemble Eq. (12) of the belief-propagation lecture notes.
3. Write down the expressions for the messages from the variables to the observations. Your expressions should look like Eq. 15-18 in the lecture notes.
4. Write down the expressions for the messages from the observations to the variables. Your expressions should look like Eq. 19-22 in the lecture notes.
5. Which path computes the marginal $p(z_1|y_1, y_2)$? Clearly show corresponding messages in the calculation of marginal from the joint. Your answer should look like the expression in page 19.