MODEL SOLUTIONS

SETTER: Neil Lawrence

Data Provided:
None

DEPARTMENT OF COMPUTER SCIENCE AUTUMN SEMESTER 2015–2016

MACHINE LEARNING AND ADAPTIVE INTELLIGENCE 15 minutes reading time and 1 hour writing time

Please note that the rubric of this paper is made different from many other papers.

Section A consists of THREE questions (Questions 1-3) and 40 marks in total.
Section B consists of TWO questions (Questions 4, 5) worth 60 marks each.

Answer ALL Questions 1-3 in Section A, and ONE Question from Section B (either Question 4 or Question 5).

Figures in square brackets indicate the percentage of available marks allocated to each part of a question.

SECTION A

Answer all three questions in this section.

1. **Multiple Choice Questions: choose EXACTLY ONE answer to each part**

PeepingTim.co are a film rental site that charges a monthly fee for membership and provides users with ranked lists of films that they recommend are next viewed. PeepingTim.co uses *matrix factorization* to assess their users preferences and the subject matter of the films. PeepingTim.co ask their users to rate films they have viewed and they have a large data base of ratings, $y_{i,j}$, where user i rates film j with a score from 1 to 5. The presence or absence of a particular rating of a particular film in the data base is represented by $s_{i,j}$ with $s_{i,j} = 1$ if user i has rated film j in our data base and $s_{i,j} = 0$ otherwise. PeepingTim.co pre-process their data by subtracting off the data mean μ from the ratings.

a) PeepingTim.co decides to minimize a sum of squares error to approximate the factors of the matrix. If user i has a taste preference represented by a vector $\mathbf{u}_i$ and the subject of a film is represented by a vector $\mathbf{v}_j$, which of the following would mathematically represent the objective function. [3%]

(i) $\sum_{i,j} s_{i,j}(\mathbf{u}_i^\top \mathbf{v}_j)^2 - s_{i,j}y_{i,j}^2$

(ii) $\sum_{i,j} s_{i,j}(\mathbf{u}_i^\top \mathbf{v}_j - y_{i,j})^2$

(iii) $\text{Tr}\mathbf{YS} - \mathbf{UV}^\top$

(iv) $(\mathbf{y} - \mathbf{UV}^\top)^\top(\mathbf{y} - \mathbf{UV}^\top)$

ANSWER:

(ii) $\sum_{i,j} s_{i,j}(\mathbf{u}_i^\top \mathbf{v}_j - y_{i,j})^2$

b) Which of the following characteristics is *not* true of applying stochastic gradient descent to this problem. [3%]

(i) Care needs to be taken when when setting the learning rate as if it is too large the minimum can be overshot.

(ii) Updates only need to be carried out if the rating has been made.

(iii) Convergence is always guaranteed to the global optima with fewer iterations than batch gradient methods.

(iv) The use of momentum terms can speed up convergence through providing a moving average estimate of the gradient over time.

ANSWER:

(iii) Convergence is always guaranteed to the global optima with fewer iterations than batch gradient methods.

c) The recommender system is fully trained. User Simon3477 has preferences represented by the vector $\mathbf{u}_i$ with elements given by

$$\mathbf{u}_i = \begin{bmatrix} 1.4 & -1.3 & 1 \end{bmatrix}.$$

The latest "James Regent" Film, Commodore is represented by the vector

$$\mathbf{v}_j = \begin{bmatrix} 1.5 & 1 & -0.7 \end{bmatrix}$$

and the mean of the data is given by 2.1. PeepingTim.co wishes to use their matrix factorization to predict Simon3477's interest in the new film. Which rating below comes closest to that prediction.

[3%]

(i) 3.5 stars

(ii) 2 stars

(iii) 4 stars

(iv) 1 star

ANSWER:

(ii) 2 stars

d) The probabilistic interpretation of the least squares objective means that predictions from PeepingTim.co's system are approximating: [3%]

(i) The logarithm of the odds ratio between the highly rated and lowly rated films

(ii) The probability that a user will watch the film.

(iii) The mean of a Gaussian random variable representing the distribution of quality scores.

(iv) The variance of the scores for those users in the neighbourhood.

ANSWER:

(iii) The mean of a Gaussian random variable representing the distribution of quality scores.

e) PeepingTim.co's data scientist suggests principal component analysis (PCA) as an alternative approach to modelling. Which one of the following is *not* true of PCA? [3%]

(i) The the principal components can be found by eigendecomposition of the data covariance matrix.

(ii) The probabilistic interpretation of PCA assumes a Gaussian prior over latent features.

(iii) The likelihood in probabilistic PCA assumes that data points are conditionally independent given the model parameters.

(iv) PCA can only be applied to data which is known to be drawn from a multivariate Gaussian.

ANSWER:

(iv) PCA can only be applied to data which is known to be drawn from a multivariate Gaussian.

2. In the answer booklet, state whether each of the following mathematical equalities is True or False.

a) $p(y|x) = p(x|y)p(y)$ [2%]

ANSWER:

False

b) $p(x|y) = \frac{p(y|x)p(x)}{p(y)}$ [2%]

ANSWER:

True

c) $\frac{p(y)}{p(x)} = \frac{p(x,y)}{p(x|y)p(x)}$ [2%]

ANSWER:

True

d) $\frac{p(y,x)}{p(x)} = \frac{p(y|x)}{p(x|y)}$ [2%]

ANSWER:

False

e) $\frac{p(x,y)}{p(x)} = \frac{p(y|x)}{p(y)}$ [2%]

ANSWER:

False

3. Consider the following five equalities:

   (1) $\mathbf{Y} = \mathbf{u}\mathbf{v}^\top$

   (2) $y = \mathbf{v}^\top \mathbf{u}\mathbf{u}^\top \mathbf{v}$

   (3) $y = \mathbf{u}^\top \text{diag}(\mathbf{u})\mathbf{v}$

   (4) $y = \text{Tr}(\mathbf{u}\mathbf{v}^\top)$

   (5) $\mathbf{y} = \mathbf{v}^\top \text{diag}(\mathbf{u})$

   Find the correct equality that matches each of the following pieces of python code. Assume that the mathematical operator $\text{Tr}(\mathbf{A})$ sums the diagonal elements of the square matrix $\mathbf{A}$ and the operator $\text{diag}(\mathbf{z})$ forms a diagonal matrix with diagonal elements given by elements of $\mathbf{z}$. Assume that in python the `numpy` library as been imported as `np` and we are given `u` and `v` as one dimensional numpy arrays.

   a) `y = np.sum(u*v)` [3%]

   ANSWER:
   (4) $y = \text{Tr}(\mathbf{u}\mathbf{v}^\top)$

   b) `y = u*v` [3%]

   ANSWER:
   (5) $\mathbf{v}^\top \text{diag}(\mathbf{u})$

   c) `y = np.outer(u,v)` [3%]

   ANSWER:
   (1) $\mathbf{Y} = \mathbf{u}\mathbf{v}^\top$

   d) `y = np.sum(v*u**2)` [3%]

   ANSWER:
   (3) $y = \mathbf{u}^\top \text{diag}(\mathbf{u})\mathbf{v}$

   e) `y = np.sum(u*v)**2` [3%]

   ANSWER:
   (2) $y = \mathbf{v}^\top \mathbf{u}\mathbf{u}^\top \mathbf{v}$

SECTION B

Answer **EITHER** Question 4 **OR** Question 5 in this section.

4. This question deals with regression in machine learning.

   a) What is the difference between *extrapolation* and *interpolation*? [10%]

ANSWER:

In extrapolation you are making predictions beyond the regime of your data, whereas in interpolation you make predictions that occur between existing data points. For example in the olympics data we used in class making a prediction about the marathon winning time for the 2016 olympics is an extrapolation, making a prediction about the marathon winning time for a hypothetical 1944 olympics would be an interpolation.

   b) In a regression problem we are given a vector of real-valued targets, $\mathbf{y}$, consisting of n observations $y_1 \dots y_n$ which are associated with multidimensional inputs $\mathbf{x}_1 \dots \mathbf{x}_n$. We assume a linear relationship between y_i and $\mathbf{x}_i$ where the data are corrupted by independent Gaussian noise giving a likelihood function of the form

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2\right).$$

   where $\mathbf{X}$ is a design matrix containing all the data points, each data point being a column vector constructed by taking a row from $\mathbf{X}$.

      (i) Show that the maximum likelihood solution for the regression weights is given through solution of the following matrix equation

$$\mathbf{X}^\top \mathbf{X}\mathbf{w} = \mathbf{X}^\top \mathbf{y}.$$

[35%]

      (ii) In practice what challenges can arise when solving this equation and how could they be addressed? [15%]

ANSWER:

(i)

Here we need to maximize the likelihood with respect to the parameters $\mathbf{w}$. The negative log likelihood can be written as

$$-\log p(\mathbf{w}|\mathbf{y},\mathbf{x},\sigma^2) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{w}^\top\mathbf{x})^2 + \text{const}$$
$$= \frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 + \frac{1}{2}\mathbf{w}^\top\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\right)\mathbf{w} + \mathbf{w}^\top\frac{\sum_{i=1}^{n}(y_i\mathbf{x}_i)}{\sigma^2},$$

where the constant represents terms which don't include $\mathbf{w}$. We can now use the equalities:

$$\mathbf{X}^\top\mathbf{X} = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top$$

and

$$\mathbf{X}^\top\mathbf{y} = \sum_{i=1}^{n}\mathbf{x}_i y_i$$

and substitute in to obtain

$$-\log p(\mathbf{w}|\mathbf{y},\mathbf{x},\sigma^2) = \frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 + \frac{1}{2}\mathbf{w}^\top\left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}+\right)\mathbf{w} - \mathbf{w}^\top\frac{\mathbf{X}^\top\mathbf{y}}{\sigma^2},$$

The next step is to take gradients with respect to $\mathbf{w}$,

$$\frac{\mathrm{d}L}{\mathrm{d}\mathbf{w}} = \left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}\right)\mathbf{w} - \frac{\mathbf{X}^\top\mathbf{y}}{\sigma^2},$$

these gradients are set to zero at a minimum giving the equation

$$\mathbf{0} = \left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}\right)\mathbf{w} - \frac{\mathbf{X}^\top\mathbf{y}}{\sigma^2},$$

which can be reorganised as

$$\mathbf{X}^\top\mathbf{X}\mathbf{w} = \mathbf{X}^\top\mathbf{y},$$

as required.

(ii)

Challenges arise with numerical stability in the solution of the equation, particularly for large n when $\mathbf{x}^\top\mathbf{x}$ becomes very big and numerically unstable or $p > n$ where the matrix is low rank. The first problem is solved by not inverting the matrix directly but performing a QR decomposition, the second by moving towards regularised or Bayesian methods.

5. This question concerns dimensionality reduction. In the question you will need to use linear algebra and your understanding of latent variable models to compute the marginal likelihood and posterior density associated with probabilistic PCA.

   a) We are often presented with data containing many thousands of features. Dimensionality reduction attempts to represent each data point, $\mathbf{y}_i$, with a *latent variable*, $\mathbf{x}_i$. Describe what is meant by the term latent variable. [15%]

ANSWER:

A latent variable is a variable that is unobserved and therefore must be inferred. This is often done through placing a prior distribution over the latent variable and marginalizing it out from the model. The latent variable can be queried by computing its posterior.

   b) Consider the following Gaussian likelihood of the ith data point given its corresponding latent variable,

$$p(\mathbf{y}_i|\mathbf{W},\sigma^2,\mathbf{x}_i) = \frac{1}{(2\pi\sigma^2)^{\frac{p}{2}}}\exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{p}\left(y_{i,j}-\mathbf{w}_k^\top\mathbf{x}_i\right)^2\right),$$

   where $\mathbf{w}_k$ is a vector taken from the kth row of the *mapping matrix*, $\mathbf{W}$, which defines the linear relationship between the latent variables and σ^2 is a noise variance parameter. Given a Gaussian prior distribution for the q dimensional latent variables,

$$p(\mathbf{x}_i) = \frac{1}{(2\pi)^{\frac{q}{2}}}\exp\left(-\frac{1}{2}\mathbf{x}_i^\top\mathbf{x}_i\right),$$

   show that the posterior distribution for a single latent variable is given by

$$p(\mathbf{x}_i|\mathbf{W},\sigma^2,\mathbf{y}_i) = \frac{1}{(2\pi)^{\frac{q}{2}}|\Sigma_x|}\exp\left(-\frac{1}{2}\left(\mathbf{x}_i-\mu_x\right)^\top\Sigma^{-1}\left(\mathbf{x}_i-\mu_x\right)\right),$$

   where q is the dimensionality of the latent space and the covariance and mean of the posterior are given as $\Sigma_x = \left(\sigma^{-2}\mathbf{W}^\top\mathbf{W}+\mathbf{I}\right)^{-1}$ and $\mu_x = \sigma^{-2}\Sigma_x\mathbf{W}^\top\mathbf{y}_i$. [30%]

ANSWER:

Need to marginalize the latent variables. But for Gaussian we use the following trick.

$$p(\mathbf{y}_i|\mathbf{W},\sigma^2)p(\mathbf{x}_i|\mathbf{y}_i) = p(\mathbf{y}_i|\mathbf{W},\sigma^2,\mathbf{x}_i)p(\mathbf{x}_i)$$

First rewrite the likelihood exponent:

$$\sum_{j=1}^{p}\left(y_{i,j}-\mathbf{w}_k^\top\mathbf{x}_i\right)^2 = \mathbf{y}_i^\top\mathbf{y}_i - 2\mathbf{y}_i^\top\mathbf{W}\mathbf{x}_i + \mathbf{x}_i^\top\mathbf{W}^\top\mathbf{W}\mathbf{x}_i$$

Combine with the prior term to give:

$$\log p(\mathbf{x}_i,\mathbf{y}_i) = -\frac{1}{2\sigma^2}\mathbf{y}_i^\top\mathbf{y}_i + \sigma^{-2}\mathbf{y}_i^\top\mathbf{W}\mathbf{x}_i - \frac{1}{2}\mathbf{x}_i^\top\left(\sigma^{-2}\mathbf{W}^\top\mathbf{W}+\mathbf{I}\right)\mathbf{x}_i + \text{const}$$

want to marginalize out $\mathbf{x}_i$ so obtain quadratic form of Gaussian in $\mathbf{x}_i$. Define $\Sigma_x = \left(\sigma^{-2}\mathbf{W}^\top\mathbf{W}+\mathbf{I}\right)^{-1}$ and $\mu_x = \Sigma_x\sigma^{-2}\mathbf{W}^\top\mathbf{y}_i$

$$\log p(\mathbf{x}_i,\mathbf{y}_i) = -\frac{1}{2\sigma^2}\mathbf{y}_i^\top\mathbf{y}_i - \frac{1}{2}(\mathbf{x}_i-\mu_x)^\top\Sigma_x^{-1}(\mathbf{x}_i-\mu_x) + \frac{1}{2}\mu_x^\top\Sigma_x^{-1}\mu_x + \text{const}$$

which now has the form of a posterior multiplied by a marginal likelihood.

c) Use the properties of multivariate Gaussians and the fact that our likelihoods imply that

$$\mathbf{x}_i \sim \mathcal{N}(\mathbf{0},\mathbf{I})$$

$$\varepsilon_i \sim \mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})$$

$$\mathbf{y}_i = \mathbf{W}\mathbf{x}_i + \varepsilon_i$$

to show that the form of the marginal likelihood for any given data point isgiven by

$$p(\mathbf{y}_i|\mathbf{W},\sigma^2) = \frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{C}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\mathbf{y}_i^\top\mathbf{C}^{-1}\mathbf{y}_i\right),$$

where the covariance matrix of this Gaussian is given by $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$. [15%]

ANSWER:

Properties of multivariate Gaussians tell us that if

$$\mathbf{x} = \mathbf{W}\mathbf{z}$$

and

$$\mathbf{z} \sim \mathcal{N}(\mu,\Sigma)$$

then

$$\mathbf{z} \sim \mathcal{N}(\mathbf{W}\mu,\mathbf{W}\Sigma\mathbf{W}^\top)$$

. Further, if two Gaussian variables are added then the result is also Gaussian with a mean that is the sum of the means and a covariance which is the sum of the covariances. The first rule tells us that

$$\mathbf{W}\mathbf{x}_i \sim \mathcal{N}(\mathbf{0},\mathbf{W}\mathbf{W}^\top)$$

combining with the second rule gives us

$$\mathbf{W}\mathbf{x}_i + \varepsilon_i \sim \mathcal{N}(\mathbf{0},\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})$$

as required.

END OF QUESTION PAPER