Continuous Optimization in Data Science
Assignment 1

Ashkan Panahi

Deadline Thursday November 9 at 23:59. Total points:100.

Q1. Gaussian Models. A real random variable X is Gaussian and denoted by $x \sim \mathcal{N}(\mu, \sigma^2)$ if its pdf is given by

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{1}$$

where μ is the mean of x and σ^2 is its variance, that is $\mathbb{E}[X] = \mu$ and $\text{Var}[X] = \sigma^2$. More generally, a multivariate Gaussian random vector $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{R})$ is a collection of random variables $\mathbf{x} = (X_1, X_2, \dots, X_n)$ given by the following joint pdf

$$p(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi\mathbf{R})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{R}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{2}$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2 \dots, \mu_n)$ is the mean vector with $\mu_i = \mathbb{E}[X_i]$ and $\mathbf{R} = (R_{ij})$ is its covariance matrix, i.e $R_{i,j} = \text{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$.

1. Consider a dataset $\mathbf{X} = \{x_1, x_2, \dots, x_n\}$ of the grades of n students in a test. Each x_i is the grade of a student (a scalar). We want to know how reliable the exam was. For this, we assume that $x_i \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d, we want to learn parameters μ, σ.

   (a) Write the negative log-likelihood function $-\log L(\mu, \sigma^2; \mathbf{X})$ (3p)

   (b) Find the maximum likelihood solution. (7p)

   (c) For a test in a classroom, discuss what μ, σ^2 might mean and what might be a suitable value of them (large or small). (5p)

2. In this part we consider linear regression. Take a dataset $\{(y_i \in \mathbb{R}, \mathbf{x}_i \in \mathbb{R}^m)\}_{i=1}^n$. We would like to understand how the variable y_i depends on $\mathbf{x}_i$. In particular, we consider the house pricing data set including the price per unit area (last column) as y and six variables $\mathbf{x} \in \mathbb{R}^6$, including the date (in decimal) of transaction, age and location of the house.

   (a) In linear regression, we assume that $y_i = \boldsymbol{\theta}^T \mathbf{x}_i + \mu + \nu_i$ where $\nu_i \sim \mathcal{N}(0, \sigma^2)$ are independent of each other and $\boldsymbol{\theta}, \mu$ are fixed but unknown. We would like to learn $\boldsymbol{\theta}, \mu, \boldsymbol{\sigma}$. Suppose that there is a prior distribution $p(\boldsymbol{\theta}) = \mathcal{N}(0, \lambda\mathbf{I})$ for $\boldsymbol{\theta}$, where $\mathbf{I}$ is the identity matrix. Write a likelihood function as the conditional distribution $p(\{y_i\} \mid \{\mathbf{x}_i\}, \boldsymbol{\theta}, \mu, \sigma^2)$.(5p)

(b) Show that the maximum likelihood estimator of $\boldsymbol{\theta}, \mu, \sigma$ corresponds to an ordinary least squares (OLS) problem (7p):

$$\hat{\boldsymbol{\theta}}, \hat{\mu} = \text{argmin}_{\boldsymbol{\theta}, \mu} \sum_{i=1}^{n} (y_i - \boldsymbol{\theta}^T \mathbf{x}_i - \mu)^2 \tag{3}$$

and

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\boldsymbol{\theta}}^T \mathbf{x}_i - \hat{\mu})^2 \tag{4}$$

(c) Show that the MAP estimator is the so-called ridge-regression, i.e. a least squares term with an ℓ_2 regularization (3p):

$$\hat{\boldsymbol{\theta}}, \hat{\mu} = \text{argmin}_{\boldsymbol{\theta}, \mu} \sum_{i=1}^{n} (y_i - \boldsymbol{\theta}^T \mathbf{x}_i - \mu)^2 + \lambda \|\boldsymbol{\theta}\|_2^2 \tag{5}$$

what is the estimate of σ in this case?

(d) Implement a program that applies OLS and ridge regression to the house pricing data set. The script pricing.py has few suggestions for your implementation. In particular, you may use the sklearn library to implement the models. Submit your code together with a short report including the following (10p):

i. What do we learn from the estimate $\hat{\boldsymbol{\theta}}$? Can you for example say which variable is more important in the pricing?

ii. Is the intercept μ important? What happens if we ignore it?

iii. what does $\hat{\sigma}$ mean? Is it better for it to be small or big?

iv. For ridge regression, how does the coefficient $\hat{\boldsymbol{\theta}}$ change with increasing λ? how does $\hat{\sigma}$ change with it?

v. Divide the dataset to the two parts related to the transactions of 2012 and 2013. Apply the models to each part separately. Do you see any difference between the two years? Is the individual models per year are more reliable than one model?

3. In the general multivariate Gaussian model the covariance $\mathbf{R}$ reflects the relation of different variables $\{X_i\}$. The inverse of the covariance matrix $\mathbf{S} = \mathbf{R}^{-1}$ is called the precision matrix. In practice, the relation between variables is visualized by using the precision matrix as a (weighted) graph. For this reason, the techniques that learn $\mathbf{R}$ (and hence $\mathbf{S}$) are sometimes referred to as graph learning (or graphical) methods. In this exercises, we consider a data set $\{\mathbf{x}_i\}$ where each sample $\mathbf{x}_i = (X_1^i, X_2^i, \ldots, X_m^i) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{R})$ is i.i.d. In our dataset, each sample is a patient diagnosed with breast cancer and each component X_k^i is the amount of activity of a particular gene k in their DNA. We want to learn how the activity of each gene may affect the activity of other genes.

(a) Write the likelihood function as the joint pdf $p(\{\mathbf{x}_i\} \mid \mathbf{R}, \boldsymbol{\mu})$ (5p).

(b) Consider a prior $p(\mathbf{R}) \propto e^{-\lambda \mathrm{Tr}[\mathbf{R}^{-1}]} = e^{-\lambda \mathrm{Tr}[\mathbf{S}]}$ where Tr(.) the trace operator. Show that the MAP estimator is given by the sample mean and regularized sample covariance matrices (8p):

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{n} \mathbf{x}_i}{n} \tag{6}$$

and

$$\hat{\mathbf{R}} = \frac{\sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T}{n} + \lambda \mathbf{I} \tag{7}$$

4. Now apply the simple graphical model in part (b) to the cancer patient's dataset. We would like to calculate the sample mean and covariance. Note that there are more than 60000 Genes in this dataset, which makes it difficult to calculate and visualize the entire covariance matrix. Instead consider two cases: a) pick the first 100 genes. b) calculate the sample variance of all genes and pick the largest 100 values. Calculate the sample covariance matrix for the chosen genes (you can use numpy matrix operations to simplify this calculation) and its inverse, i.e. the sample precision matrix $\hat{\mathbf{S}}$. Visualize the result as a graph. For this, choose a threshold t and assume an edge only when the corresponding element of $\hat{\mathbf{S}}$ is larer tan t in absolute value. You can use the Networkx library to create this graph and visualize it. there is a python script file (gene.py) to help you. Make a report of the result including (17p):

   (a) What is the effect of λ. Can you set it to zero? why not? when can you?

   (b) How you select the threshold.

   (c) comparing the two cases of different selected genes.

   (d) Include the results for a number of different thresholds

   (e) Take a look at the corresponding mean values $\hat{\boldsymbol{\mu}}$. What does it imply to have a big mean value?

Q2. Logistic Regression

1. Now, we consider a dataset $\{(y_i \in \{0,1\}, \mathbf{x}_i \in \mathbb{R}^n)\}$. We want to find a relation between $\mathbf{y}_i$ and $\mathbf{x}_i$, but $\mathbf{y}_i$ is binary. The example that we consider is again related to breast cancer. $\mathbf{x}$ is a list of activity level of 10 different genes and $y = 0$ means no cancer is detected, while $y = 1$ means cancer is detected. For a bianry variable $Y \in \{0,1\}$, we define its odds as the following ratio:

$$\mathrm{odds}(Y) = \frac{\Pr[Y = 1]}{\Pr[Y = 0]} \tag{8}$$

In logistic regression, we assume

$$\text{odds}(Y \mid \mathbf{x}) = e^{\boldsymbol{\theta}^T \mathbf{x} + \mu} \tag{9}$$

where $\boldsymbol{\theta}, \mu$ are the model parameters. Write the negative log likelihood function for the dataset. (Hint: First, you need to calculate $\Pr[Y \mid \mathbf{x}]$ from the odds) (5p).

2. In the next modules of the course, we will see that by the first part of this assignment, the maximum likelihood problem is convex and hence can be efficiently solved. You can use its implementation in the sklearn library (look at cancer.py for more instructions). When loading the data, note that variable 1 is much bigger than the others. Apply logistic regression to the first 500 samples and learn $\boldsymbol{\theta}, \mu$. Consider two cases: In the first one, take variable 1 as is. In the second one, divide it by 10^6 in every sample. Then, calculate the odds for the remaining 183 samples. For every test sample with odds larger than 1, we predict that $y = 1$ and otherwise $y = 0$. Compare it to the actual labels and count how many predictions are correct. This is known as accuracy. Write a short report about the result (10p):

   (a) Does the second case (i.e. preprocessing the data) give you a different solution? Why do you think?

   (b) What can we learn from $\boldsymbol{\theta}$? is there any particular gene that has a bigger effect on cancer?

   (c) How well is the accuracy? Discuss what is the reason and what are the practical considerations for using the learned model.

Q3.

1. Consider the LP minimal form:

$$\begin{aligned} &\min_{\mathbf{x} \in \mathbb{R}^n} \mathbf{c}^T \mathbf{x} \\ &\text{s.t.} \\ &\mathbf{A}\mathbf{x} + \mathbf{b} \leq \mathbf{0} \end{aligned} \tag{10}$$

   By repeating the argument in the classroom show that the dual optimization can be written in the standard form (7p):

$$\begin{aligned} &\min_{\boldsymbol{\lambda} \in \mathbb{R}^n} \mathbf{b}^T \boldsymbol{\lambda} \\ &\text{s.t.} \\ &\boldsymbol{\lambda} \geq \mathbf{0} \\ &\mathbf{A}^T \boldsymbol{\lambda} + \mathbf{c} = \mathbf{0} \end{aligned} \tag{11}$$

2. Consider the set $S = \{(y, x) \in \mathbb{R}^2 \mid y \geq |x|\}$.

   (a) Show that S is convex (4p).

   (b) Find every supporting vecor (i.e. all normal vectors) of S at $(0, 0)$ (4p).