In the Name of God

Introduction to Machine Learning (25737-2)

Problem Set 01

Spring Semester 1401-02

Department of Electrical Engineering

Sharif University of Technology

Instructor: Dr. S. Amini

Due on Esfand 12, 1401 at 23:55

($*$) starred problems are optional and have a bonus mark!

Some Formulas to Use

- Suppose $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)^\top$ is a jointly Gaussian random vector with the following parameters:

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \qquad \boldsymbol{\Delta} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Delta}_{11} & \boldsymbol{\Delta}_{12} \\ \boldsymbol{\Delta}_{21} & \boldsymbol{\Delta}_{22} \end{pmatrix}.$$

Then the marginal distributions are given by:

$$\begin{aligned} p_{\mathbf{X}_1}(\mathbf{x}_1) &= \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}), \\ p_{\mathbf{X}_2}(\mathbf{x}_2) &= \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}). \end{aligned} \tag{1}$$

And the posterior distribution is given by:

$$\begin{aligned} p_{\mathbf{X}_1|\mathbf{X}_2}(\mathbf{x}_1|\mathbf{x}_2) &= \mathcal{N}(\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}) \\ \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2) = -\boldsymbol{\mu}_1 - \boldsymbol{\Delta}_{11}^{-1}\boldsymbol{\Delta}_{12}(\mathbf{y}_2 - \boldsymbol{\mu}_2) \\ \boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Delta}_{11}^{-1} \end{aligned} \tag{2}$$

- $$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}}$$

1 Correlation, Causality, and Independence

Let $X \sim \mathrm{Uniform}(-1,1)$, and $Y = X^2$. Clearly, X and Y aren't independent. (Actually, they have a causation property!). Show that even though they are dependant, they are uncorrelated, which means $\rho_{X,Y} = 0$.

2 Markov-Chain Gaussians

We write $X \to Y \to Z$ and say that $X, Y,$ and Z form a Markov chain when we have: $X|Y \perp\!\!\!\perp Z|Y$ which also means $p_{X,Z|Y}(x,z|y) = p_{X|Y}(x|y)p_{Z|Y}(z|y)$. For three Gaussians variables with the preceding property, compute $\rho_{X,Z}$ in terms of $\rho_{X,Y}$ and $\rho_{Y,Z}$.

3 Sensor Fusion

Imagine the temperature is a fixed number z (which we know nothing about. You can model it with $Z \sim \mathcal{N}(0, +\infty)$). We have two sensors, in which the temperature is measured with noise. The variance of noise for each of them is known and it's v_1 and v_2 respectively. Suppose we make n_1 observation from the first sensor, each given by $\{Y_1^{(i)}\}_{i=1}^{n_1}$ and n_2 observation of the second sensor given by $\{Y_2^{(i)}\}_{i=1}^{n_2}$. Consider all of these observations to be shown as a set called $\mathcal{D}$. Using the given variances, find $p_{Z|\mathcal{D}}(z|\mathcal{D})$ and estimate Z using its mean.

Hint: assume that Z is a Normal random variable with a prior and try to estimate the posterior using the equations given at the beginning of the assignment. First, you have to formalize $\{Y_1^{(i)}\}_{i=1}^{n_1}, \{Y_2^{(i)}\}_{i=1}^{n_2}$ and their parameters in an accurate way.

4 Pseudo Inverse

Assume that matrix $\mathbf{A}$ has an SVD decomposition as follows $\mathbf{A} = \mathbf{U\Sigma}V^\top$. We define the pseudo-inverse of $\mathbf{A}$ as $\mathbf{A}^\dagger = \mathbf{V\Sigma}^{-1}\mathbf{U}^\top$. Prove the followings:

1. if $\mathbf{A}$ has a full row rank, then $\mathbf{A}^\dagger = \mathbf{A}^\top(\mathbf{AA}^\top)^{-1}$.
2. if $\mathbf{A}$ has a full column rank, then $\mathbf{A}^\dagger = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top$.

5 Eigenvalues

We show the eigenvalues of the square matrix $\mathbf{A}$ by $\lambda_1, \lambda_2, \ldots, \lambda_n$. Prove the followings:

1. $$\mathrm{Tr}\,\{\mathbf{A}\} = \sum_{i=1}^{n} \lambda_i.$$
2. $$\det\{\mathbf{A}\} = \prod_{i=1}^{n} \lambda_i.$$

6 Maximum Likelihood Estimation

Suppose we have a random vector $\mathbf{X} \in \mathbb{R}^d$. All elements are assumed to be iid random variables. Assume that we have an observation $\mathbf{x}$. We want to fit a probability distribution to this data and we are going to use the maximum likelihood for that.

1. Assume that each X_i is a Bernouli random variable, i.e., $p_{X_i}(x_i) = \theta^{x_i}(1-\theta)^{1-x_i}$. Also assume that we have observed m ones and k zeros. Find the distribution parameter θ.
2. Assume that each X_i is an Exponential random variable, i.e., $p_{X_i}(x_i) = \lambda e^{-\lambda x_i}\mathbf{1}\{x_i \geq 0\}$. Also assume that all x_i values are positive. Find the exponential parameter λ.
3. Assume that each X_i is a Normal random variable, i.e., $p_{X_i}(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$. Find the mean and variance of the distribution.

7 A Tiny Bit of Vector Differentiation

Prove the following differentiation formulas. These formulas will be useful throughout the course.

1.
$$\nabla_{\mathbf{x}} \left(\mathbf{a}^\top \mathbf{x}\right) = \left[\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_d}\right]^\top \left(\mathbf{a}^\top \mathbf{x}\right) = \mathbf{a}^\top.$$

2.
$$\nabla_{\mathbf{x}} \left(\mathbf{x}^\top \mathbf{A}\mathbf{x}\right) = \left[\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_d}\right]^\top \left(\mathbf{x}^\top \mathbf{A}\mathbf{x}\right) = x^\top(\mathbf{A} + \mathbf{A}^\top) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}.$$

8 (∗) Bayes Rule for Gaussian Variables

Prove the equation (2).

9 (∗) Implicit Regularization!

Assume that $\mathbf{x} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{m\times n}$, and $\mathbf{b} \in \mathbb{R}^m$. We are trying to solve the following problem:

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$$

1. Assume that $m < n$, prove that in this case, there are multiple answers for the problem. Also show that $\mathbf{x}^* = \mathbf{A}^\dagger \mathbf{b}$ is an answer for tha minimization problem and has the smallest l_2 norm among all answers.

2. Now let's to solve the optimization problem using gradient descent method. Assume we start from $\mathbf{x}^{(0)} = \mathbf{0}$ and use this formula for updates (ϵ is the learning rate):

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \epsilon \mathbf{A}^\top (\mathbf{A}\mathbf{x}^{(t)} - \mathbf{b})$$

. Prove that by choosing an appropriate value for learning rate, after a sufficient number of iterations, $\mathbf{x}^{(t)}$ converges to $\mathbf{x}^*$.

3. Show that $\frac{2}{\sigma^2_{\max}(\mathbf{A})}$ is a good upper bound for the learning rate. This means that for any $\epsilon \le \frac{2}{\sigma^2_{\max}(\mathbf{A})}$, if we use the gradient descent method, the objective function will monotonically decrease. Which means:
$$\|\mathbf{A}\mathbf{x}^{(t+1)} - \mathbf{b}\|_2 \le \|\mathbf{A}\mathbf{x}^{(t)} - \mathbf{b}\|_2.$$