HW1 Handwritten Assignment

Sep 2023

Problem 1 Preliminary (1 pt)

(a) (0.2 pts) Given $\mathbf{w} \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{m \times m}$. Show that

$$\nabla_{\mathbf{w}} \mathbf{w}^\mathsf{T} \mathbf{A} \mathbf{w} = \mathbf{A}^\mathsf{T} \mathbf{w} + \mathbf{A} \mathbf{w}.$$

In particular, if $\mathbf{A}$ is a symmetric matrix, then

$$\nabla_{\mathbf{w}} \mathbf{w}^\mathsf{T} \mathbf{A} \mathbf{w} = 2\mathbf{A} \mathbf{w}$$

(b) (0.2 pts) Given $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times m}$. Show that

$$\frac{\partial \ tr(\mathbf{AB})}{\partial a_{ij}} = b_{ji} \tag{1}$$

where

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mm} \end{bmatrix}$$

It is common to write (1) as

$$\frac{\partial \ tr(\mathbf{AB})}{\partial \mathbf{A}} = \mathbf{B}^\mathsf{T}.$$

(c) (0.6 pts) Prove that

$$\frac{\partial \log(\det(\mathbf{A}))}{\partial a_{ij}} = \mathbf{e}_j^\mathsf{T} \mathbf{A}^{-1} \mathbf{e}_i, \tag{2}$$

where $\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix} \in \mathbb{R}^{m \times m}$ is a (non-singular) matrix, and $\mathbf{e}_j$ is the unit vector along the j-th axis (e.g. $\mathbf{e}_3 = [0, 0, 1, 0, ..., 0]^T$). It is common to write (2) as

$$\frac{\partial \log(\det(\mathbf{A}))}{\partial \mathbf{A}} = \left(\mathbf{A}^{-1}\right)^\mathsf{T}$$

Problem 2 Classification using Gaussian (2.4 pts)

In this question, we tackle the binary classification problem through the generative approach, where we assume the data point X (viewed as a $\mathbb{R}^d$-valued r.v.) and its label Y (viewed as a $\{\mathcal{C}_1, \mathcal{C}_2\}$-valued r.v.) are generated according to the generative model (paramerized by θ) as follows:

$$\mathbb{P}_\theta[X = \mathbf{x}, Y = \mathcal{C}_k] = \pi_k f_{\boldsymbol{\mu}_k, \Sigma_k}(\mathbf{x}) \quad (k \in \{1, 2\}) \tag{3}$$

where $\theta = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$ for which

$$f_{\boldsymbol{\mu}_k, \Sigma_k}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\intercal \Sigma_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\right).$$

Now suppose we observe data points $\mathbf{x}_1, ..., \mathbf{x}_N$ and their corresponding labels $y_1, ..., y_N$.

(a) (1.2 pt)

  (i) (0.3 pt) Please write down the likelihood function $L(\theta)$ that describes how likely the generative model would generate the observed data $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ in terms of $\theta = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$.

  (ii) (0.3 pt) Find the maximum likelihood estimate $\theta^* = (\pi_1^*, \pi_2^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*, \Sigma_1^*, \Sigma_2^*)$ that maximizes the likelihood function $L(\theta)$.

  (iii) (0.3 pt) Write down $\mathbb{P}_\theta[Y = \mathcal{C}_1 | X = \mathbf{x}]$ and $\mathbb{P}_\theta[X = \mathbf{x} | Y = \mathcal{C}_1]$ in terms of $\theta = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$. What are the physical meaning of the aforementioned quantities?

  (iv) (0.3 pt) Express $\mathbb{P}_\theta[X = \mathbf{x} | Y = \mathcal{C}_1]$ in the form of $\sigma(z)$, where $\sigma(\cdot)$ denotes the sigmoid function, and express z in terms of $\theta = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$ and x.

(b) (1.2 pt) Suppose we pose an additional constraint that the covariance matrices of the two Gaussian distributions are identical, namely $\Sigma_1 = \Sigma_2 = \Sigma$, in which the generative model is parameterized by $\vartheta = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$. Redo questions (a) under such setting.

Problem 3 (Application of Gaussian classifier) (0.6 pts)

In this question, you will train a binary classifier based on the data which can be downloaded from https://reurl.cc/2EZMzn following the settings in Problem 2. Each data point and its label take the format $x_i \in \mathbb{R}^2$ and $y_i \in \{0, 1\}$.

(a) (0.2 pts) Calculate $\vartheta^* = (\pi_1^*, \pi_2^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*, \Sigma^*)$ as in Problem 2 (b) in numbers.

(b) (0.2 pts) Calculate $\theta^* = (\pi_1^*, \pi_2^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*, \Sigma_1^*, \Sigma_2^*)$ as in Problem 2 (a)(ii) in numbers.

(c) (0.2 pts) Please draw the scatter plot of the data. Which model is better in your opinion between (a) and (b)? Why?

Problem 4 (Closed-Form Linear Regression Solution) (1 pts + Bonus 1.5 pts)

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

where $\mathbf{y} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n\times d}, \boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{\epsilon} \in \mathbb{R}^n$. Denote $\mathbf{X}_i \in \mathbb{R}^{1\times d}$ as the i-th row of $\mathbf{X}$, with the following interpretations:

- If the linear model has the bias term, then write $\boldsymbol{\theta} = [w_1, \cdots, w_m, b]^\mathsf{T}$ and $\mathbf{X}_i = [x_{i,1}, x_{i,2}, \cdots, x_{i,m}, 1]$, namely $d = m + 1$.
- If the linear model has no bias term, then write $\boldsymbol{\theta} = [w_1, \cdots, w_d]^T$ and $\mathbf{X}_i = [x_{i,1}, x_{i,2}, \cdots, x_{i,m}]$, namely $d = m$.

(a) Without the bias term, consider the L^2-regularized loss function:

$$\sum_i \kappa_i \left(y_i - \boldsymbol{X}_\mathrm{i}\boldsymbol{\theta}\right)^2 + \lambda \sum_j w_j^2,\ \lambda > 0.$$

Show that the optimal solution that minimizes the loss function is $\boldsymbol{\theta^*} = \left(\boldsymbol{X}^T\boldsymbol{K}\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{K}\boldsymbol{y}$, where

$$\boldsymbol{K} = \begin{bmatrix} \kappa_1 & & 0 \\ & \ddots & \\ 0 & & \kappa_n \end{bmatrix}$$

is a diagonal matrix and $\boldsymbol{I}$ is the $n \times n$ identical matrix.

(b) (Bonus, 1.5 pts) With the bias term, the L^2-regularized loss function becomes

$$\sum_i \kappa_i \left(y_i - \boldsymbol{X}_\mathrm{i}\boldsymbol{\theta} - b\right)^2 + \lambda \sum_j w_j^2,\ \lambda > 0.$$

Show that the optimal solution that minimizes the loss function is $\boldsymbol{\theta^*} = [\boldsymbol{w}^{\star T}, b^\star]^T$, where

$$\boldsymbol{w}^\star = \left(\tilde{\boldsymbol{X}}^T\boldsymbol{K}\tilde{\boldsymbol{X}} + \lambda\boldsymbol{I} - \tilde{\boldsymbol{X}}^T\boldsymbol{K}\boldsymbol{e}\boldsymbol{e}^T\boldsymbol{K}\tilde{\boldsymbol{X}}\right)^T \tilde{\boldsymbol{X}}^T\boldsymbol{K}\left(\boldsymbol{y} - \boldsymbol{e}^T\boldsymbol{y}\boldsymbol{e}\right),$$

$$b^\star = \left(\boldsymbol{e}^T\boldsymbol{y} - \boldsymbol{e}^T\boldsymbol{K}\tilde{\boldsymbol{X}}\boldsymbol{w}\right)^T$$

for which $\boldsymbol{e} = [1\ ...\ 1]^T$ denotes the all one vector, $\boldsymbol{X} = [\tilde{\boldsymbol{X}}\boldsymbol{e}]$, and that $\mathbf{K}$ and $\mathbf{I}$ are defined as in (a).

Problem 5 (Noise and Regulation) (1 pts)

Consider the linear model $f_{\mathbf{w},b} : \mathbb{R}^k \to \mathbb{R}$, where $\mathbf{w} \in \mathbb{R}^k$ and $b \in \mathbb{R}$, defined as

$$f_{\mathbf{w},b}(x) = \mathbf{w}^T\mathbf{x} + b$$

Given dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, if the inputs $\mathbf{x}_i \in \mathbb{R}^k$ are contaminated with input noise $\boldsymbol{\eta}_i \in \mathbb{R}^k$, we may consider the expected sum-of-squares loss in the presence of input noise as

$$\tilde{L}_{ss}(\mathbf{w}, b) = \mathbb{E}\left[\frac{1}{2N}\sum_{i=1}^{N}\left(f_{\mathbf{w},b}(\mathbf{x}_i + \boldsymbol{\eta}_i) - y_i\right)^2\right]$$

where the expectation is taken over the randomness of input noises $\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_N$. Additionally, the inputs $(\mathbf{x}_i)$ and the input noise $(\boldsymbol{\eta}_i)$ are independent.

Now assume the input noises $\eta_i = [\eta_{i,1}\ \eta_{i,2}\ ...\eta_{i,k}]$ are random vectors with zero mean $\mathbb{E}[\eta_{i,j}] = 0$, and the covariance between components is given by

$$\mathbb{E}[\eta_{i,j}\eta_{i',j'}] = \delta_{i,i'}\delta_{j,j'}\sigma^2$$

where $\delta_{i,i'} = \begin{cases} 1 & \text{, if } i = i' \\ 0 & \text{, otherwise.} \end{cases}$ denotes the Kronecker delta.

Please show that

$$\tilde{L}_{ss}(\mathbf{w}, b) = \frac{1}{2N}\sum_{i=1}^{N}\left(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i\right)^2 + \frac{\sigma^2}{2}\|\mathbf{w}\|^2$$

That is, minimizing the expected sum-of-squares loss in the presence of input noise is equivalent to minimizing noise-free sum-of-squares loss with the addition of a L^2-regularization term on the weights. (Hint: $\|\mathbf{x}\|^2 = \mathbf{x}^\mathbf{T}\mathbf{x} = \mathbf{tr}(\mathbf{x}\mathbf{x}^\mathbf{T})$ and the square of a vector is dot product with itself)

Problem 6 (Mathematical Background) (0 pt)

Please click the following link

`https://www.cs.cmu.edu/~mgormley/courses/10601/homework/hw1.zip`

to download the Homework 1 from CMU 2023 Machine Learning Website. You could finish Section 3 to Section 6 of this homework. If you have any problem, please ask TA at TA hour.

Some Tools You Need to Know

1. Orthogonal Matrix
2. Positive Definite, Semipositive Definite

3. Eigenvalue Decomposition, Singular value decomposition
4. Lagrange Multiplier
5. Trace

You can find the definition and the usage by yourself. It is also welcome to discuss with TA in TA hour.

As next homework. Give intermediate subproblems to hint how to progress

Problem XX (Gradient Descent Convergence)

Suppose the function $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, and that its gradient is Lipschitz continuous with constant $L > 0$, i.e. we have that $||\nabla f(x) - \nabla f(y)||_2 \leq L||x - y||_2$ for any x, y. Then if we run gradient descent for k iterations with a fixed step size $t \leq \frac{1}{L}$, it will yield a solution $x^{(k)}$ which satisfies

$$f(x^{(k)}) - f(x^\star) \leq \frac{||x^{(0)} - x^*||_2}{2tk}$$

where $f(x^\star)$ is the optimal value.