Boston University

ENG EC 414 Introduction to Machine Learning

Exam 1 Solution

Released on Monday, 5 October, 2020 (120 minutes, 42 points + 2 bonus points), submit to Gradescope

---

- *There are 6 problems plus 1 bonus one.*
- *For each part, you must clearly outline the key steps and provide proper justification for your calculations in order to receive full credit.*
- *You can use any material from the class (slides, discussions, homework solutions, etc.), but you cannot look for solutions on the internet. Also, be aware of the limited time.*

Problem 1.1 [5pts] Let $f(z) := z^2$ and $\mathcal{A} := [-1, 1]$. Compute: $\operatorname{argmin}_{z \in \mathcal{A}} \frac{1}{13+\sqrt{1+2 \cdot f(z)}}$.

Solution:

5pts $\boxed{\operatorname{argmin}_{z \in \mathcal{A}} \frac{1}{13+\sqrt{1+2 \cdot z^2}} = \{-1, +1\}}$ First of all, the minimum of the function is equivalent to the maximum of $13 + \sqrt{1+2 \cdot f(z)}$, because $1/x$ is decreasing on the positive values of x. In turn, the function $g(t) = 13 + \sqrt{(1 + 2t)}$ is strictly increasing for all $t > 0$ since its derivative $1/\sqrt{1+2t}$ is strictly positive $\Rightarrow \operatorname{argmax}_{z \in \mathcal{A}} \left[13 + \sqrt{1+2 \cdot f(z)}\right] = \operatorname{argmax}_{z \in \mathcal{A}} g(f(z)) = \operatorname{argmax}_{z \in \mathcal{A}} f(z)$. The quadratic z^2 decreases from 1 at $z = -1$ to 0 at $z = 0$ and then increases to 1 at $z = 1$.

Problem 1.2 [6pts] Let $\mathcal{Y} := \{1.5, \ 2.0, \ 3.5, \ 6.0\}$ and $y_1 = y_2 = 1.5, \ y_3 = 2.0, \ y_4 = y_5 = 3.5, \ y_6 = 6.0$.

(a) [2pts] Compute $\operatorname{argmin}_{y \in \mathcal{Y}} \frac{1}{6} \sum_{j=1}^{6} \mathbf{1}[y \neq y_j]$.

(b) [2pts] Compute $\operatorname{argmin}_{y \in \mathbb{R}} \frac{1}{6} \sum_{j=1}^{6} (y - y_j)^2$.

(c) [2pts] Compute $\operatorname{argmin}_{y \in \mathbb{R}} \frac{1}{6} \sum_{j=1}^{6} |y - y_j|$.

Solution: Note that the values 1.5 and 3.5 occur twice and the values 2 and 6 each occur once in the dataset.

(a) [2pts] $\boxed{\{1.5, 3.5\}}$. The average 0-1 loss is minimized by the most frequent values. These are 1.5 and 3.5.

(b) [2pts] $\boxed{\{3.0\}}$. The mean square error is minimized by the empirical mean which is $\frac{2\times1.5+1\times2+2\times3.5+1\times6}{6} = 3.0$.

(c) [2pts] $\boxed{[2, 3.5]}$. The mean absolute error is minimized by the empirical median which is any point in the closed (continuous) interval $[2, 3.5]$.

Problem 1.3 [10pts]

(a) [2pts] Consider the hyperplane parametrized by $\mathbf{w}$ and b with $b = 3$ and $\mathbf{w} = (1,\ -4,\ 8)^\top$. Determine which of the following points lie on the hyperplane: (i) $\mathbf{x}_1 = (-2,\ 2,\ 1)^\top$, (ii) $\mathbf{x}_2 = (0,\ 1,\ 0)^\top$, (iii) $\mathbf{x}_3 = (1,\ 3,\ 1)^\top$.

(b) [2pts] Compute the distance of $\mathbf{x}_4 = (-1,\ -1,\ -1)^\top$ from the hyperplane in part (a).

(c) [3pts] Compute the orthogonal projection of the point $\mathbf{x}_4$ from part (b) onto the hyperplane in part (a).

(d) [3pts] Determine parameters $\mathbf{w}$ and b of the hyperplane passing through the following 3 points: $\mathbf{x}_5 = (1/2,\ 0,\ 0)^\top$, $\mathbf{x}_6 = (1,\ 1,\ 0)^\top$, $\mathbf{x}_7 = (-1,\ 1,\ -1)^\top$.

Solution:

(a) [2pts] $\boxed{\mathbf{x}_3}$ A point $\mathbf{x}$ lies on $hp(\mathbf{w}, b) \Leftrightarrow \mathbf{w}^\top\mathbf{x} + b = 0$. We have $\mathbf{w}^\top\mathbf{x}_1 + b = +1$, $\mathbf{w}^\top\mathbf{x}_2 + b = -1$, $\mathbf{w}^\top\mathbf{x}_3 + b = 0$.

(b) [2pts] $\boxed{2/9}$ Distance of point $\mathbf{x}_4$ from $hp(\mathbf{w}, b) = \frac{|\mathbf{w}^\top\mathbf{x}_4 + b|}{\|\mathbf{w}\|} = \frac{|-1+4-8+3|}{\sqrt{1^2+(-4)^2+(8)^2}} = \frac{2}{9}$.

(c) [3pts] $\boxed{-(79/81,\ 18/81,\ 65/81)^\top}$ $\mathrm{Proj}_{hp(\mathbf{w},b)}(\mathbf{x}_4) = \mathbf{x}_4 - \frac{(\mathbf{w}^\top\mathbf{x}_4 + b)}{\|\mathbf{w}\|} \cdot \frac{\mathbf{w}}{\|\mathbf{w}\|} = \mathbf{x}_4 - \frac{(-2)}{9}\frac{\mathbf{w}}{9} = -(79/81,\ 89/81,\ 65/81)^\top$.

(d) [3pts] $\boxed{\mathbf{w} = b \cdot (-2, 1, 4)^\top,\ \text{any } b \neq 0}$ Points $\mathbf{x}_5, \mathbf{x}_6, \mathbf{x}_7$ lie on $hp(\mathbf{w}, b) \Rightarrow \mathbf{w}^\top\mathbf{x}_j + b = 0$, $j = 5, 6, 7$. This gives 3 linear equations in 4 unknowns:

$$\begin{bmatrix} 1/2 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ -1 & 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad \Rightarrow \quad \begin{aligned} \tfrac{1}{2}w_1 + b &= 0 \\ w_1 + w_2 + b &= 0 \\ -w_1 + w_2 - w_3 + b &= 0 \end{aligned} \quad \begin{aligned} \therefore w_1 &= -2b, \\ \therefore w_2 &= -w_1 - b = b, \\ \therefore w_3 &= -w_1 + w_2 + b = 4b. \end{aligned}$$

Choose any $b \neq 0$.

Problem 1.4 [6pts] Consider the following set of feature vectors and corresponding real-valued labels

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} -3 \\ 1 \end{bmatrix}, \mathbf{x}_4 = \begin{bmatrix} 4 \\ 3 \end{bmatrix}, \quad y_1 = 4, y_2 = 2, y_3 = -8, y_4 = 2\ .$$

(a) [4pts] Fix $b = 0$ and compute by hand the ordinary least squares (OLS) solution $\mathbf{w}^\star$.

(b) [2pts] Compute the OLS prediction of $\mathbf{w}^\star$ and $b = 0$ for the vector $\mathbf{x}_1$.

Solution:

(b) [3pts]

$$\mathbf{w}^\star_{OLS} = (X^\top X)^{-1} X\mathbf{y} = 2 \times \frac{1}{13\times 4 - 3\times 3}\begin{bmatrix} 4 & -3 \\ -3 & 13 \end{bmatrix}\begin{bmatrix} 10 \\ -1 \end{bmatrix} = \frac{2}{43}\begin{bmatrix} 43 \\ -43 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$$

(c) [2pts]

$$h_{OLS}(\mathbf{x}_1) = (\mathbf{w}^\star_{OLS})^\top \mathbf{x}_1 = \begin{bmatrix} 2 & -2 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = 4$$

Problem 1.5 [7pts] Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ be a training set with feature vectors $\mathbf{x}_j \in \mathbb{R}^d$ and labels $y_j \in \mathbb{R}$. Consider the following cost function for Regularized Least Square without bias, that is, there is no b:

$$g(\mathbf{w}) = \|\mathbf{w}\|^2 + \frac{1}{2m}\sum_{j=1}^{m}(y_j - \mathbf{x}_j^\top \mathbf{w})^2 .$$

<u>Note that this formulation is slightly different from the one seen in class, don't just copy from the slides!</u>

(a) [2pts] Compute the gradient $\nabla g(\mathbf{w})$.

(b) [2pt] Provide pseudocode for an algorithm to minimize $g(\mathbf{w})$ based on gradient descent with zero initialization, a fixed positive step size $\eta > 0$, and the maximum number of iterations T.

(c) [3pt] After a certain number of iterations less than the maximum number of iterations, $\boldsymbol{w}_t$ in gradient descent stops changing, that is $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t$. Can it happen? If yes, in which situations? If no, why?

Solution:

(a) [2pts]

$$\nabla g(\mathbf{w}) = 2\mathbf{w} + \frac{1}{m}\sum_{j=1}^{m}(\mathbf{x}_j^\top \mathbf{w} - y_j)\mathbf{x}_j$$

Instead, in matrix form and using the usual notation on X, it would be

$$\nabla g(\mathbf{w}) = 2\mathbf{w} + \frac{1}{m}X^\top X\mathbf{w} - \frac{1}{m}X^\top \boldsymbol{y}$$

Both are good.

(b) [1pt]

Pseudocode

input: $\boldsymbol{y}, X, \eta, T$
initialize: $\mathbf{w}_1 = \mathbf{0}$
for $t = 1, 2, \ldots, T$
  compute gradient $\nabla g(\mathbf{w}_t)$**:** $\mathbf{v}_t = 2\mathbf{w}_t + \frac{1}{m}\sum_{j=1}^{m}(\mathbf{x}_j^\top \mathbf{w}_t - y_j)\mathbf{x}_j$
  update w: $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{v}_t$
end for
output: w

(c) [3pts] If gradient descent stops updating, the only possibility is that the gradient is exactly zero. For a convex function, this implies we are exactly in the minimum.

Problem 1.6 [8pts] Consider the following training set of feature vectors and corresponding binary labels

$$\mathbf{x}_1 = \begin{bmatrix}1\\2\end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix}3\\1\end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix}-3\\1\end{bmatrix}, \mathbf{x}_4 = \begin{bmatrix}1\\3\end{bmatrix}, \quad y_1 = -1, y_2 = 1, y_3 = 1, y_4 = -1 \ .$$

(a) [2pts] Hand-plot the training set. Proper labeling of axes and key points is needed to receive full credit.

(b) [2pts] Is it possible to find a hyperplane that linearly separates this training set? A motivation for your answer is needed to receive full credit.

(c) [2pts] Will the Perceptron converge on this dataset? A motivation for your answer is needed to receive full credit.

(d) [2pts] Using the usual augmentation to include the bias in features and hyperplane, compute by hand the first update $\tilde{\boldsymbol{w}}_2$ of the Perceptron algorithm starting from $\tilde{\boldsymbol{w}}_1 = [0, 0, 0]^\top$, after seeing the example $\tilde{\mathbf{x}}_1 = \begin{bmatrix}1\\\mathbf{x}_1\end{bmatrix}$.

Solution:

(a) [2pts]

(b) [2pts] The dataset is linearly separable. For example, the sign of predictions of the hyperplane $\mathbf{w} = \begin{bmatrix}0\\-1\end{bmatrix}$, $b = 1.5$ correctly classify all the samples.

(c) [2pts] Given that the dataset is linearly separable with a positive margin, the Perceptron will converge. Note that the Perceptron will converge not matter what is the order of the examples.

(d) [2pts]

$$\tilde{\boldsymbol{w}}_2 = \tilde{\boldsymbol{w}}_1 + y_1\tilde{\mathbf{x}}_1 = \begin{bmatrix}-1\\-1\\-2\end{bmatrix}$$

Problem 1.7 [Bonus, 2pts] Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ equal to $f(x_1, x_2) = \frac{1}{2}x_1^2 + \frac{1}{5}x_2^2 + \frac{1}{4}\sin(2x_1)$. Is it convex? Motivate your answer.

Solution: First, don't be fooled by the presence of the sin function. To estabilish convexity, we have to calculate the Hessian. We can calculate the Hessian of this function easily, to get $H(x_1, x_2) = \begin{bmatrix}1 - \sin(2x_1) & 0\\ 0 & \frac{2}{5}\end{bmatrix}$. This matrix is PSD: there are at least a couple of ways to see it. First method: for a diagonal matrix, the eigenvalues are equal to the element of the diagonal, that are non-negative for any (x_1, x_2). Second method: use the definition of PSD, so we have to show that $\mathbf{z}^\top H(x_1, x_2)\mathbf{z} \geq 0$ for any $\mathbf{z}, x_1, x_2$. This is true because $\mathbf{z}^\top H(x_1, x_2)\mathbf{z} = z_1^2(1 - \sin(2x_1)) + \frac{2}{5}z_2^2$. Hence, the function is convex.