Assignment II

Machine Learning II
2021-2022 - UMONS
Souhaib Ben Taieb

1

Solve Problem 1.12 in LFD.

Problem 1.12 This problem investigates how changing the error measure can change the result of the learning process. You have N data points $y_1 \leq \cdots \leq y_N$ and wish to estimate a 'representative' value.

(a) If your algorithm is to find the hypothesis h that minimizes the in sample sum of squared deviations,

$$E_{\text{in}}(h) = \sum_{n=1}^{N} (h - y_n)^2,$$

then show that your estimate will be the in sample mean,

$$h_{\text{mean}} = \frac{1}{N} \sum_{n=1}^{N} y_n.$$

(b) If your algorithm is to find the hypothesis h that minimizes the in sample sum of absolute deviations,

$$E_{\text{in}}(h) = \sum_{n=1}^{N} |h - y_n|,$$

then show that your estimate will be the in sample median h_{med}, which is any value for which half the data points are at most h_{med} and half the data points are at least h_{med}.

(c) Suppose y_N is perturbed to $y_N + \epsilon$, where $\epsilon \to \infty$. So, the single data point y_N becomes an outlier. What happens to your two estimators h_{mean} and h_{med}?

Figure 1: Source: Abu-Mostafa et al. Learning from data. AMLbook.

2

Solve Problem 3.12 in LFD.

Problem 3.12 In linear regression, the in sample predictions are given by $\hat{\mathbf{y}} = \mathrm{H}\mathbf{y}$, where $\mathrm{H} = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}$. Show that H is a projection matrix, i.e. $\mathrm{H}^2 = \mathrm{H}$. So $\hat{\mathbf{y}}$ is the projection of $\mathbf{y}$ onto some space. What is this space?

Figure 2: Source: Abu-Mostafa et al. Learning from data. AMLbook.

3

Solve Problem 3.17 in LFD.

Problem 3.17 Consider a function

$$E(u,v) = e^u + e^{2v} + e^{uv} + u^2 - 3uv + 4v^2 - 3u - 5v,$$

(a) Approximate $E(u+\Delta u, v+\Delta v)$ by $\hat{E}_1(\Delta u, \Delta v)$, where $\hat{E}_1$ is the first-order Taylor's expansion of E around $(u,v) = (0,0)$. Suppose $\hat{E}_1(\Delta u, \Delta v) = a_u \Delta u + a_v \Delta v + a$. What are the values of a_u, a_v, and a?

(b) Minimize $\hat{E}_1$ over all possible $(\Delta u, \Delta v)$ such that $\|(\Delta u, \Delta v)\| = 0.5$. In this chapter, we proved that the optimal column vector $\begin{bmatrix}\Delta u \\ \Delta v\end{bmatrix}$ is parallel to the column vector $-\nabla E(u,v)$, which is called the *negative gradient direction*. Compute the optimal $(\Delta u, \Delta v)$ and the resulting $E(u+\Delta u, v+\Delta v)$.

(c) Approximate $E(u+\Delta u, v+\Delta v)$ by $\hat{E}_2(\Delta u, \Delta v)$, where $\hat{E}_2$ is the second order Taylor's expansion of E around $(u,v) = (0,0)$. Suppose

$$\hat{E}_2(\Delta u, \Delta v) = b_{uu}(\Delta u)^2 + b_{vv}(\Delta v)^2 + b_{uv}(\Delta u)(\Delta v) + b_u \Delta u + b_v \Delta v + b.$$

What are the values of b_{uu}, b_{vv}, b_{uv}, b_u, b_v, and b?

(d) Minimize $\hat{E}_2$ over all possible $(\Delta u, \Delta v)$ (regardless of length). Use the fact that $\nabla^2 E(u,v)\big|_{(0,0)}$ (the Hessian matrix at $(0,0)$) is positive definite to prove that the optimal column vector

$$\begin{bmatrix}\Delta u^* \\ \Delta v^*\end{bmatrix} = -\left(\nabla^2 E(u,v)\right)^{-1} \nabla E(u,v),$$

which is called the *Newton direction*.

(e) Numerically compute the following values:

(i) the vector $(\Delta u, \Delta v)$ of length 0.5 along the Newton direction, and the resulting $E(u+\Delta u, v+\Delta v)$.

(ii) the vector $(\Delta u, \Delta v)$ of length 0.5 that minimizes $E(u+\Delta u, v+\Delta v)$, and the resulting $E(u+\Delta u, v+\Delta v)$. (*Hint: Let* $\Delta u = 0.5 \sin\theta$.)

Compare the values of $E(u+\Delta u, v+\Delta v)$ in (b), (e i), and (e ii). Briefly state your findings.

The negative gradient direction and the Newton direction are quite fundamental for designing optimization algorithms. It is important to understand these directions and put them in your toolbox for designing learning algorithms.

Figure 3: Source: Abu-Mostafa et al. Learning from data. AMLbook.

4

In this exercise, we ask you to prove that gradient descent (GD) converges to the global optimum for L-**Lipschitz convex** functions, which are a rich class of functions that cover many problems in machine learning.

Definition 1. $f : \mathbb{R}^d \to \mathbb{R}$ is **convex** if for all $x, y \in \mathbb{R}^d$ and $0 \leq \lambda \leq 1$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y).$$

Definition 2. A smooth function $f : \mathbb{R}^d \to \mathbb{R}$ is L-**Lipschitz** if for all $x, y \in \mathbb{R}^d$,

$$\left\|\nabla f(x) - \nabla f(y)\right\|_2 \leq L\|x-y\|_2 .$$

Property 1. If $f : \mathbb{R}^d \to \mathbb{R}$ is **convex**, then for all x, y,

$$f(x) + \langle \nabla f(x), y - x \rangle \leq f(y), \tag{1}$$

where $\langle \cdot, \cdot \rangle$ is the dot product.

Property 2. If $f : \mathbb{R}^d \to \mathbb{R}$ is L-**Lipschitz convex**, then for all x, y,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y-x\|_2^2 . \tag{2}$$

In the following, gradient descent (GD) will refer to the following algorithm:

1. Choose $x_0 \in \mathbb{R}^d$ and step-size $\eta > 0$
2. For $i = 0, 1, 2, \ldots$, compute
$$x_{i+1} = x_i - \eta \nabla f(x_i)$$

Theorem. If $f : \mathbb{R}^d \to \mathbb{R}$ is L-**Lipschitz convex**,, and $x^* = \operatorname{argmin}_x f(x)$, then GD with step-size $\eta \leq \frac{1}{L}$ satisfies:

$$f(x_k) \leq f(x^*) + \frac{\|x_0 - x^*\|}{2\eta k}. \tag{3}$$

First, show that

$$f(x_{i+1}) \leq f(x_i) - \frac{\eta}{2}\left\|\nabla f(x_i)\right\|_2^2 \tag{4}$$

for $i = 0, 1, \ldots, k$. To do so, apply the following steps:

1. Use the fact that f is L-**Lipschitz convex**, i.e. use (2) with $y = x_{i+1}$ and $x = x_i$.
2. Replace x_{i+1} by $x_i - \eta \nabla f(x_i)$, and use the fact that $\langle a, a \rangle = \|a\|_2^2$.
3. Use the fact that $L\eta \leq 1$.

Secondly, show that

$$f(x_{i+1}) \leq f(x^*) + \frac{1}{2\eta}\left(\left\|x_i - x^*\right\|_2^2 - \left\|x_{i+1} - x^*\right\|_2^2\right) \tag{5}$$

To do so, apply the following steps:

1. Combine (4) with the fact that f is convex, i.e. by (1), we have
$$f(x_i) \leq f(x^*) + \langle \nabla f(x_i), x_i - x^* \rangle.$$

2. Use the fact that $\nabla f(x_i) = \frac{1}{\eta}(x_i - x_{i+1})$.

3. Show that

$$f(x_{i+1}) \leq f(x^*) + \frac{1}{2\eta}\left\|x_i - x^*\right\|_2^2 - \frac{1}{2\eta}\left\|x_i - x^* - \eta\nabla f(x_i)\right\|_2^2.$$

Note that

$$\frac{1}{\eta}\langle a, b\rangle - \frac{1}{2\eta}\|a\|_2^2 = -\frac{1}{2\eta}\left(-2\langle a, b\rangle + \|a\|_2^2\right)$$

can be rewritten, by completing the square, as follows

$$\frac{1}{2\eta}\|b\|_2^2 - \frac{1}{2\eta}\left(\|b\|_2^2 - 2\langle a, b\rangle + \|a\|_2^2\right) = \frac{1}{2\eta}\|b\|_2^2 - \frac{1}{2\eta}\|b - a\|_2^2.$$

Finally, by summing (5) for $i = 0, \ldots, k-1$, show that

$$\sum_{i=0}^{k-1}(f(x_{i+1}) - f(x^*)) \leq \frac{1}{2\eta}\left\|x_0 - x^*\right\|_2^2, \tag{6}$$

Furthermore, since by (4), $f(x_0), f(x_1), \ldots, f(x_k)$ is non-increasing, we have $f(x_k) - f(x^*) \leq f(x_i) - f(x^*)$ for all $i < k$. Thus, we have

$$k(f(x_k) - f(x^*)) \leq \frac{1}{2\eta}\left\|x_0 - x^*\right\|_2^2,$$

which proves (3).

TURN IN

- A jupyter notebook with your solutions.
- **DUE**: March 27, 11:55pm (late submissions not allowed), loaded into Moodle.