Machine Learning, 2021 Fall
Assignment 1

Notice

Due 23:59 (CST), Oct. 23, 2021

Plagiarizer will get 0 points.
LaTeXis highly recommended. Otherwise you should write as legibly as possible.

1 Gradient Descent

In order to minimize $f(\boldsymbol{x})$ where $x \in R^n$, we takes iteration:

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \alpha_k \boldsymbol{p}^k$$

where $\boldsymbol{p}^k = \boldsymbol{H}^k \nabla f\left(\boldsymbol{x}^k\right)$ and $\alpha^k \to 0^+$. What kind of $\boldsymbol{H}^k$ can guarantee that $\boldsymbol{p}^k$ is a descent direction ? Give a detailed proof.[1pts]

Solution:
In order to let the objective descent, we should have:

$$f(x^{k+1}) = f(x^k + \alpha_k p^k) = f(x^k + \alpha_k H^k \nabla f(x^k)) < f(x^k)$$

Take Taylor expansion at x_k:

$$f(x^{k+1}) = f(x^k) + \nabla f(x^k)^T \alpha_k H^k \nabla f(x^k) + o(\alpha_k H^k \nabla f(x^k))$$

Thus:

$$\nabla f(x^k)^T \alpha_k H^k \nabla f(x^k) + o(\alpha_k H^k \nabla f(x^k)) < 0$$

As $\lim_{\alpha_k \to 0^+} o(\alpha_k H^k \nabla f(x^k)) = 0$:

$$f(x^k)^T H^k \nabla f(x^k) < 0$$

2 Convex

(1) Prove that $f : R^n \to R$ is a convex function if and only if $epif = \{(x,t) \in R^{n+1} | x \in dom(f),\ f(x) \leq t\}$ is a convex set. [0.5pts]
(2) Let $f_1, f_2, ..., f_k$ be convex functions on R^n, prove that $f(x) = max\{f_1(x), f_2(x), ..., f_k(x)\}$ is also a convex function. [0.5pts]
(3) Prove that $f : R^n \to R$ is a convex function if and only if $g : R \to R$

$$g(t) = f(x + tv) \quad dom(g) = \{t | x + tv \in dom(f)\}$$

is convex for any $x \in dom(f)$ and $v \in R^n$. [0.5pts]
(4) Prove that $f(x) = log(det(x)) \quad dom(f) = S^n_{++}$ is a concave function. [0.5pts]

Solution:
(1):
Sufficiency:
Suppose f is convex and $(x_1, t_1), (x_2, t_2) \in epif$. $\theta \in [0, 1]$

$$f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta)f(x_2) \leq \theta t_1 + (1-\theta)t_2$$

thus $(\theta x_1 + (1-\theta)x_2, \theta t_1 + (1-\theta)t_2) \in epif$ Necessity:
Suppose $epif$ is a convex set.

$$\forall (x_1, f(x_1)), (x_2, f(x_2)) \in epif,\ \theta \in [0,1] \quad (\theta x_1 + (1-\theta)x_2, \theta f(x_1) + (1-\theta)f(x_1)) \in epif$$

thus:

$$\forall x_1, x_2 \in domf \quad f(\theta x_1 + (1-\theta)x_2)) \leq \theta f(x_1) + (1-\theta)f(x_2)$$

(2):

$$epif = \bigcap_{i=1}^{k} epif_i$$

Thus $epif$ is the intersection of convex sets, it is also a convex set, which implies f is a convex function.
(3):
Sufficiency:
Suppose f is a convex function. $t_1, t_2 \in domg$, $\theta \in [0, 1]$

$$\begin{aligned} g(\theta t_1 + (1-\theta)t_2) &= f(x + \theta t_1 v + (1-\theta)t_2 v) \\ &= f(\theta(x + t_1 v) + (1-\theta)(x + t_2 v)) \\ &\leq \theta f(x + t_1 v) + (1-\theta) f(x + t_2 v) \\ &= \theta g(t_1) + (1-\theta) g(t_2) \end{aligned}$$

Thus g is convex.
Necessity:
Suppose $g(t)$ is convex, $x_1, x_2 \in domf$.
Let $g(t) = f(x_1 + t(x_2 - x_1))$

$$f(\theta x_1 + (1-\theta)x_2) = g(\theta 0 + (1-\theta)1) \leq \theta g(0) + (1-\theta)g(1) = \theta f(x_1) + (1-\theta)f(x_2)$$

Thus f is convex.
(4):
Let $g(t) = logdet(x + tv)$

$$\begin{aligned} g(t) &= logdet(x + tv) \\ &= logdet(x^{\frac{1}{2}}(I + tx^{-\frac{1}{2}} V x^{-\frac{1}{2}})x^{\frac{1}{2}}) \\ &= logdet(x) + logdet(I + tx^{-\frac{1}{2}} V x^{-\frac{1}{2}}) \end{aligned}$$

Let $Z = I + tx^{-\frac{1}{2}}Vx^{-\frac{1}{2}} = I + tQ\begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{bmatrix}Q^T = Q\begin{bmatrix}1+t\lambda_1 & & \\ & \ddots & \\ & & 1+t\lambda_n\end{bmatrix}Q^T$

$$g(t) = logdet(x) + \sum_{i=1}^{n} log(1 + t\lambda_i)$$

thus $-g(t)$ is convex, $-f$ is also convex by (3), then f is concave.

3 Learning

Assume that $\mathcal{X} = \{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_N, \mathrm{x}_{N+1}, \ldots, \mathrm{x}_{N+M}\}$, $N, M \in \mathbb{N}^+$ and $\mathcal{Y} = \{-1, +1\}$ with an unknown target function $f : \mathcal{X} \to \mathcal{Y}$. The training data set $\mathcal{D}$ is $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$. Define the off-training-set error of a hypothesis h with respect to f by

$$E_{\text{off}}(h, f) = \frac{1}{M}\sum_{m=1}^{M}\left[h\left(\mathbf{x}_{N+m}\right) \neq f\left(\mathbf{x}_{N+m}\right)\right]$$

(a) Say $f(\mathrm{x}) = +1$ for all x and

$$h(\mathbf{x}) = \begin{cases} +1, & \text{for } \mathbf{x} = \mathbf{x}_k \text{ and } k \text{ is odd and } 1 \leq k \leq M+N \\ -1, & \text{otherwise} \end{cases}$$

What is $E_{\text{off}}\left(h, f\right)$? [0.5pts]

(b) We say that a target function f can 'generate' $\mathcal{D}$ in a noiseless setting if $y_n = f\left(\mathbf{x}_n\right)$ for all $\left(\mathbf{x}_n, y_n\right) \in \mathcal{D}$. For a fixed $\mathcal{D}$ of size N , how many possible $f : \mathcal{X} \to \mathcal{Y}$ can generate $\mathcal{D}$ in a noiseless setting? [0.25pts]

(c) For a given hypothesis h and an integer k between 0 and M , how many of those f in (b) satisfy $E_{\text{off}}\left(h, f\right) = \frac{k}{M}$? [0.25pts]

(d) For a given hypothesis h , if all those f that generate $\mathcal{D}$ in a noiseless setting are equally likely in probability, what is the expected off trainingset error $\mathbb{E}_f\left[E_{\text{off}}(h, f)\right]$? [0.5pts]

(e) A deterministic algorithm A is defined as a procedure that takes $\mathcal{D}$ as an input, and outputs a hypothesis $h = A(\mathcal{D})$. Argue that for any two deterministic algorithms A_1 and A_2. [0.5pts]

$$\mathbb{E}_f\left[E_{\text{off}}\left(A_1(\mathcal{D}), f\right)\right] = \mathbb{E}_f\left[E_{\text{off}}\left(A_2(\mathcal{D}), f\right)\right]$$

Solution:

(a) $E_{\text{off}}\left(h, f\right)$ equals to the fraction of even numbers between $N+1$ and $N+M$ in total M numbers.

(b) For a function f to generate $\mathcal{D}$ in a noiseless setting, it means $f\left(x_i\right) = y_i$ for $i = 1, 2, \ldots, N$. They only have freedom to assign various values on the rest points in $\mathcal{X}$, i.e. $x_{N+1}, \ldots, x_{N+M}$. So there are total $2^M f$ that can generate $\mathcal{D}$.

(c) Among those f in(b), if they agree with a given h on $M-k$ points, i.e. $E_{off}(h, f) = \frac{k}{M}$, then each of them need to choose k points from M $x_{N+1}, \ldots, x_{N+M}$ to match with y , and other M-k points to mismatch with y . This has $\begin{pmatrix} M \\ k \end{pmatrix}$ combinations.

(d) Given a hypothesis h , for an integer k between 0 and M , from previous problem (c), we know that there are $\begin{pmatrix} M \\ k \end{pmatrix}$ number of functions f that satisfy $E_{off}(h, f) = \frac{k}{M}$. So the probability to get $E_{off}(h, f) = \frac{k}{M}$ is $\frac{\begin{pmatrix} M \\ k \end{pmatrix}}{2^M}$. So the expectation is $E_f\left[E_{off}(h, f)\right] = \frac{1}{2^M}\sum_{k=0}^{M}\begin{pmatrix} M \\ k \end{pmatrix}\frac{k}{M} = \frac{1}{2}$.

(e) The above expectation in problem (d) depends on M only, and doesn't depend on h at all (each term in the expectation depends on the number of mismatches between h and f , but they are consumed in the expectation). So for any two deterministic algorithms A_1 and A_2 , the expectations will be the same.

4 MAE

The Empirical risk minimization(ERM) principle is meant to choose a hypothesis $\hat{h}$ which minimizes the empirical risk $\hat{R}_{\mathcal{D}}[h]$.

(a) Consider the following hypothesis and loss function

$$\mathcal{H} = \{h_\theta(x) = \theta_1 x : \theta_1 \in \mathbb{R}\},$$
$$\mathcal{L}(\theta_1) = \frac{1}{N}\sum_{i=1}^{N}\left|h_\theta\left(x^{(i)}\right) - y^{(i)}\right|$$

Assume we already have a dataset $\mathcal{D} = \{(1,3),(-1,-2),(2,4)\}$. Derive the value of θ_1 which minimizes the empirical risk. [0.5pts]

(b) Assume we have a dataset $\mathcal{D} = \{x_1, x_2, \cdots, x_n\}$ where $x_i \in \mathcal{R}$. Consider the hypothesis $\mathcal{H}$ to be

$$\mathcal{H} = \{h_\theta = \theta_0 : \theta_0 \in \mathbb{R}\}$$

Derive the hypothesis h^* which minimizes the empirical risk. [0.5pts]

$$h^* = \arg\min_h \frac{1}{N}\sum_{i=1}^{N}\left|h - x^{(i)}\right| \quad \text{s.t.} \quad h \in \mathcal{H}$$

Solution:

(a) $\theta_1^* = 2$.

(b) Taking the derivative of the empirical risk w.r.t h and let it equal to 0, we have

$$\sum_{n=1}^{N}\text{sign}(h - x_n) = 0$$

Let $h^* = \text{median}(x_i),\ i = 1, 2, \cdots, N$ be the median of the data points. Then we have half of the data smaller than h^*, and half of the data larger than h^*, which makes the derivative to zero.