Machine Learning, 2022 Spring
Assignment 1

Problem 1 If $\mu = 0.9$, what is the probability that a sample of 10 marbles will have $\nu \leq 0.1$?

[Hints: 1. Use binomial distribution. 2. The answer is a very small number.]

By the problem, $\nu \leq 0.1 \Leftrightarrow$ a sample of 10 marbles will have at most 1 red marble, then we have

$$\begin{aligned} P(\nu \leq 0.1) &= P(1 \text{ red marble }) + P(0 \text{ red marble }) \\ &= (1-0.9)^{10} + C_1^{10} \times 0.9 \times (1-0.9)^9 \\ &= 9.1 \times 10^{-9} \end{aligned}$$

Problem 2 If $\mu = 0.9$, use the Hoeffding Inequality to bound the probability that a sample of 10 marbles will have $\nu \leq 0.1$ and compare the answer to the previous exercise.

In this problem,

$$P[|\nu - \mu| \geq 0.8] \leq 2e^{-2*0.8^2*10} = 5.52 * 10^{-6}$$

$9.1 * 10^{-9} < 5.52 * 10^{-6}$, which satisfies Hoeffding's Inequality.

Problem 3 We are given a data set $\mathcal{D}$ and of 25 training examples from an unknown target function $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \{-1, +1\}$. To learn f, we use a simple hypothesis set $\mathcal{H} = \{h_1, h_2\}$ and, where h_1 is the constant $+1$ function and h_2 is the constant -1.

We consider two learning algorithms, S (smart) and C (crazy). S chooses the hy pothesis that agrees the most with $\mathcal{D}$ and C chooses the other hypothesis deliberately. Let us see how these algorithms perform out of sample from the deterministic and probabilistic points of view. Assume in the probabilistic view that there is a probability distribution on $\mathcal{X}$, and let $\mathbb{P}[f(x) = +1] = p$.

(a) Can S produce a hypothesis that is guaranteed to perform better than random on any point outside $\mathcal{D}$?

(b) Assume for the rest of the exercise that all the examples in $\mathcal{D}$ have $y_n = +1$. Is it possible that the C hypothesis that produces turns out to be better than the hypothesis that S produces?

(c) If $p = 0.9$, what is the probability that S will produce a better hypothesis than C?

(d) Is there any value of p for which it is more likely than not that C will produce a better hypothesis than S ?

Solution:

(a) Using the S algorithm will choose the hypothesis with good effect on the training set, but it cannot guarantee the good performance in the part outside the training set.

(b) Assuming that $y_n = +1$ on the training set are all 1, the accuracy of the hypothesis obtained by using the C algorithm on the training set is 0, and the accuracy of the hypothesis obtained by using the S algorithm on the training set is 1, but the part outside the training set is unknown, so the hypothesis obtained by the C algorithm may still perform better than the hypothesis obtained by the S algorithm.

(c) Let the output of S is $f_s, theoutputofC$ is f_c, the probability that we need to figure out is

$$\mathbb{P}\left(\mathbb{P}\left[f_s = f\right] > \mathbb{P}\left[f_c = f\right]\right)$$

Since $f_s(x)$ always be $1, \mathbb{P}[f(x) = +1] = 0.9$,

$$\mathbb{P}[f_s = f] = \mathbb{P}[f(x) = +1] = p = 0.9$$
$$\mathbb{P}[f_c = f] = 1 - \mathbb{P}[f_s = f] = 0.1$$

Thus,

$$\mathbb{P}(\mathbb{P}[f_s = f] > \mathbb{P}[f_c = f]) = \mathbb{P}(0.9 > 0.1) = 1$$

(d) This is a generalization of the last problem, and they're actually asking if there is any p, such that

$$\mathbb{P}(\mathbb{P}[f_s = f] > \mathbb{P}[f_c = f]) < 0.5$$

Since $\mathbb{P}[f(x) = +1] = p$,

$$\mathbb{P}[f_s = f] = p, \mathbb{P}[f_c = f] = 1 - p$$
$$\mathbb{P}(\mathbb{P}[f_s = f] > \mathbb{P}[f_c = f]) = \mathbb{P}(p > 1 - p)$$

Thus when $p < 0.5$,

$$\mathbb{P}(\mathbb{P}[f_s = f] > \mathbb{P}[f_c = f]) = 0 < 0.5$$

Problem 4 Given functions $g(x) = \|Qx - b\|_2^2$ and $f(x) = x^T Qx + b^T x, Q \in \mathbb{R}^{n \times n}$, write down the gradients and Hessian.

Solution:

Note that $\nabla_x(x^T \mathbf{Q}x) = (\mathbf{Q}^T + \mathbf{Q})x$, $\nabla_x(\mathbf{b}^T x) = \mathbf{b}$, and $\nabla_x^2(x^T \mathbf{Q}x) = \mathbf{Q}^T + \mathbf{Q}$, we have:

$$\nabla_x g(x) = 2\mathbf{Q}^T(\mathbf{Q}x - \mathbf{b}), \quad \nabla_x^2 g(x) = 2\mathbf{Q}^T\mathbf{Q}.$$

$$\nabla_x f(x) = (\mathbf{Q}^T + \mathbf{Q})x + \mathbf{b}, \quad \nabla_x^2 f(x) = \mathbf{Q}^T + \mathbf{Q}.$$

Problem 5 For symmetric $Q \in \mathbb{R}^{n \times n}$, give the definition of positive definiteness. What is the value of λ guaranteeing the positive definiteness of $\lambda I + Q^T Q$.

Solution:

Positive Definiteness: A symmetric matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is positive definite if

$$\forall x \in \mathbb{R}^n, \ x \neq 0, \ x^T \mathbf{Q}x > 0.$$

For a given symmetric matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$, by the definition of positive definiteness, we have:

$$\begin{aligned}
\lambda\mathbf{I} + \mathbf{Q}^T\mathbf{Q} \text{ is positive definite} &\iff x^T(\lambda\mathbf{I} + \mathbf{Q}^T\mathbf{Q})x > 0, \ \forall x \neq 0 \\
&\iff \lambda > -\frac{x^T\mathbf{Q}^T\mathbf{Q}x}{x^T x}, \ \forall x \neq 0 \\
&\iff \lambda > \max_{x \neq 0} -\frac{x^T\mathbf{Q}^T\mathbf{Q}x}{x^T x} \\
&\iff \lambda > -\min_{x \neq 0} \frac{x^T\mathbf{Q}^T\mathbf{Q}x}{x^T x} \\
&\iff \lambda > -\lambda_{\min}(\mathbf{Q}^T\mathbf{Q})
\end{aligned}$$

where $\lambda_{\min}(\mathbf{Q}^T\mathbf{Q})$ is the minimum eigenvalue of $\mathbf{Q}^T\mathbf{Q}$. The last inequality of follows from Rayleigh Quotient.

If we consider all symmetric matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$, we need to take an upper bound of $-\lambda_{\min}(\mathbf{Q}^T\mathbf{Q})$, which is 0. Thus, we require that $\lambda > 0$.

Problem 6 For closed set $C \subset \mathbb{R}^n$, give the definition of convex set. For function $f : \mathbb{R}^n \to \mathbb{R}$, give the definition of convex functions.

Solution:

Convex set: C is convex if $\forall \boldsymbol{x}, \boldsymbol{y} \in C, \theta \in [0, 1]$, we have

$$\theta\boldsymbol{x} + (1 - \theta)\boldsymbol{y} \in C \tag{1}$$

Convex function: f is convex if:

1. $dom(f)$ is convex.
2. $\forall \boldsymbol{x}, \boldsymbol{y} \in dom(f)$, $\theta \in [0,1]$, we have

$$f(\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y}) \leq \theta f(\boldsymbol{x}) + (1-\theta) f(\boldsymbol{y}) \tag{2}$$

Remark It is not enough to say that f is convex if $\forall \boldsymbol{x}, \boldsymbol{y} \in dom(f)$, we have

$$f\Big(\frac{\boldsymbol{x}+\boldsymbol{y}}{2}\Big) \leq \frac{1}{2}(f(\boldsymbol{x}) + f(\boldsymbol{y})) \tag{3}$$

since this condition requires that f is continuous. See Midpoint-convex doesn't imply convex.

Problem 7 Write the first 4 terms of the Taylor expansion of $f(x) = e^x + 3x + (x-1)^2$ at $\hat{x} = 0$. Write the first two terms of the Taylor expansion of $f : \mathbb{R}^n \to \mathbb{R}$ at $\hat{\boldsymbol{x}}$.

Solution:

Expanding $f(x)$ yields:

$$f(x) = \left(1 + x + \frac{x^2}{2} + \frac{x^3}{6} + o\left(x^3\right)\right) + 3x + (x-1)^2 = 2 + 2x + \frac{3x^2}{2} + \frac{x^3}{6} + o\left(x^3\right) \tag{4}$$

Note that f is a multi-variable function, of which Taylor approximation is given by

$$f(\boldsymbol{x}) \approx f(\hat{\boldsymbol{x}}) + \nabla f(\hat{\boldsymbol{x}})^T (\boldsymbol{x} - \hat{\boldsymbol{x}}) \tag{5}$$

Problem 8 Consider two independent random variables X and Y.X is continuous with probability density function $f(x)$ for $x \in \mathbb{R}$.Y is a Bernoulli distribution with probability mass function $g(y)$ for $y \in \{0,1\}$. Write the formulation of the expectation of random $\phi(X,Y)$.

Solution:

$$E(\phi(X,Y)) = g(0)\int_{-\infty}^{\infty} \phi(x,0) f(x) dx + g(1)\int_{-\infty}^{\infty} \phi(x,1) f(x) dx \tag{6}$$