Assignment 1 - TFML

Nirjhar Nath
nirjhar@cmi.ac.in
BMC202239

Solution 1:

We have,

$$\begin{aligned}\int_0^\infty \mathbb{P}(X \geq \nu)\mathrm{d}\nu &= \int_0^\infty \int_v^\infty f(x)\mathrm{d}x\mathrm{d}v \\ &= \int_0^\infty \int_0^x f(x)\mathrm{d}v\mathrm{d}x \\ &= \int_0^\infty f(x)\cdot(x-0)\mathrm{d}x \\ &= \int_0^\infty xf(x)\mathrm{d}x \\ &= \mathbb{E}[X].\end{aligned}$$

■

Solution 2:

We define the function $f(\alpha)$ as:

$$f(\alpha) = KL^+((1-\alpha)q + \alpha p \| q).$$

Clearly,

$$f(1) = KL^+(p\|q), \quad f(0) = KL^+(q\|q) = 0.$$

Next, we compute the first and second derivatives of $f(\alpha)$. The first derivative is:

$$f'(\alpha) = (p-q)\left(\ln\frac{(p-q)\alpha + q}{q} - \ln\frac{(1-q)-(p-q)\alpha}{1-q}\right).$$

Evaluating at $\alpha = 0$ gives:

$$f'(0) = 0.$$

The second derivative is:

$$f''(\alpha) = \frac{(p-q)^2}{(q+(p-q)\alpha)((1-q)-(p-q)\alpha)}.$$

At $\alpha = 0$, we obtain:

$$f''(0) = \frac{(p-q)^2}{q(1-q)}.$$

Using the Taylor expansion of $f(\alpha)$ around $\alpha = 0$, we have:

$$f(1) = KL^+(p\|q) = f(0) + f'(0) + \frac{f''(0)}{2} + \text{higher-order terms}.$$

Since $f(0) = 0$ and $f'(0) = 0$, we get:

$$KL^+(p\|q) = \frac{f''(0)}{2} + \text{higher-order terms} = \frac{(p-q)^2}{2q(1-q)} + \text{higher-order terms}.$$

Thus, we approximate $KL^+(p\|q)$ as:

$$KL^+(p\|q) \geq \frac{(p-q)^2}{2q(1-q)}.$$

For $p \geq q \geq \frac{1}{2}$ (or $p \leq q \leq \frac{1}{2}$), this simplifies to:

$$KL^+(p\|q) \geq 2(p-q)^2.$$

which completes the proof. ■

Solution 3:

Let $\mathcal{H}$ be the set of concentric circles centered at the origin. Then, we prove that the following claim is true.

Claim: $\text{VCdim}(\mathcal{H}) = 1$.
Proof: Let $h_r \in \mathcal{H}$ denote a circle of radius r, and suppose for a point $x \in X \subseteq \mathbb{R}^2$,

$$h_r(x) = \begin{cases} 1, & \text{if } ||x|| < r, \\ 0, & \text{otherwise.} \end{cases}$$

First, we observe that any subset of size 1 can be shattered by $\mathcal{H}$. For any point $x_1 \in X$, we can choose a circle such that x_1 lies inside or outside the circle. Thus, it is possible to assign any label to a single point, meaning that $\mathcal{H}$ can shatter any subset of size 1.
Now, consider a subset $\{x_1, x_2\} \subseteq X$ of size 2. We will show that no such subset can be shattered by $\mathcal{H}$.
Case 1: Suppose $||x_1|| = ||x_2||$, i.e., both points are equidistant from the origin. In this case, any circle $h_r \in \mathcal{H}$ will assign the same label to both x_1 and x_2, as they lie on the same circle. Hence, it is impossible to have different labels for p_1 and x_2, so no subset of size 2 can be shattered in this case.
Case 2: Suppose $||x_1|| \neq ||x_2||$. Without loss of generality, assume $||x_1|| < ||x_2||$. Then, the labeling $h_r(x_1) = 1$ and $h_r(x_2) = 0$ is not possible, as any circle that contains x_2 must also contain x_1. Similarly, it is impossible to have $h_r(x_1) = 0$ and $h_r(x_2) = 1$. Thus, no subset of size 2 can be shattered in this case either.
Since no subset of size 2 can be shattered, therefore, $\text{VCdim}(\mathcal{H}) = 1$.

Since $\text{VCdim}(\mathcal{H}) = 1 < \infty$, i.e., $\mathcal{H}$ has a finite VC dimension, it follows that $\mathcal{H}$ is both PAC learnable and agnostic-PAC learnable.

The same argument applies to concentric spheres in higher dimensions, as the key property is the symmetry about the origin. For concentric spheres, no subset of size 2 can be shattered, and thus the VC dimension is also 1. Therefore, the hypothesis class of concentric spheres is both PAC learnable and agnostic-PAC learnable as well. ■

Solution 4:

We are given a concept class consisting of conjunctions of at most n Boolean literals $x_1, x_2, \ldots, x_n, \overline{x_1}, \overline{x_2}, \ldots, \overline{x_n}$, where each literal can take values in $\{0, 1\}$. Let $\mathcal{H}$ be the hypothesis class of conjunctions of at most n Boolean literals. The total number of conjunctions of these literals is given by the number of ways to select and combine the literals. Since each variable x_i can take two values, either x_i or $\overline{x_i}$, the total number of hypotheses in $\mathcal{H}$ is $|\mathcal{H}| = 2^n$. Now, we can compute the VC dimension of $\mathcal{H}$.

Claim: The VC dimension of $\mathcal{H}$ is n.
Proof: We have,

$$\text{VCdim}(\mathcal{H}) \leq \log_2 |\mathcal{H}| = n.$$

To show that the VC dimension of $\mathcal{H}$ is exactly n, consider a subset $\{x_1, x_2, \ldots, x_n\} \subseteq X$, where each element in the subset corresponds to a different literal. Since we can trivially shatter this subset of size n by appropriately selecting conjunctions of literals from $\mathcal{H}$, we conclude that $\text{VCdim}(\mathcal{H}) = n$.

Since $\text{VCdim}(\mathcal{H}) = n$, i.e., $\mathcal{H}$ has a finite VC dimension, it is both PAC learnable and agnostic PAC learnable.

Next, we determine the sample complexity for PAC learning and agnostic PAC learning. For agnostic PAC learning, the sample complexity is given by the following inequality:

$$C_1 \left(\frac{n + \log(\frac{1}{\delta})}{\epsilon^2} \right) \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \left(\frac{n + \log(\frac{1}{\delta})}{\epsilon^2} \right)$$

for some constants $C_1, C_2 > 0$.

For PAC learning, the sample complexity satisfies the following bound:

$$C_1 \left(\frac{n \log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)}{\epsilon} \right) \leq C_2 \left(\frac{n \log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)}{\epsilon} \right)$$

for some constants $C_1, C_2 > 0$.

Thus, the hypothesis class of conjunctions of Boolean literals is both PAC learnable and agnostic PAC learnable, with the sample complexity bounds as described above. ■

Solution 5:

Let A be the algorithm described in the problem, and let $R \in \mathcal{C}$ be an axis-aligned rectangle. Consider $\epsilon > 0$, $m \in \mathbb{N}$, and $\mathcal{D}$ as a distribution over $\mathbb{R}^2 \times \{0, 1\}$, such that $(X, Y) \sim \mathcal{D}$, where the label $Y = 1$ occurs with probability $1 - \eta$ when $X \in R$, and with probability η, Y is flipped to 0 (i.e., $Y = 0$ occurs due to noise). When $X \notin R$, $Y = 0$ deterministically. Let $r, t, b, l \in \mathbb{R}$ such that $R = [l, r] \times [b, t]$.

Now, let $S = \{(X_1, Y_1), \ldots, (X_m, Y_m)\} \sim \mathcal{D}^m$ be a sample of m i.i.d. points, and let $R_S = A(S)$ be the rectangle learned by the algorithm A from the sample S.

If $\mathcal{P}(X \in R) < \epsilon$, then since $R_S \subseteq R$, the algorithm achieves the desired accuracy because the error can only come from $R \setminus R_S \cap R$. Hence, if the probability mass of points inside R is smaller than ϵ, we are done.

If $\mathcal{P}(X \in R) \geq \epsilon$, define four regions by slicing R into four chunks. Define $r_1 = [l, g_1] \times [b, t]$, where $g_1 = \inf\{g \geq l \mid \mathcal{P}(X \in [l, g] \times [b, t]) \geq \epsilon/4\}$, which takes a "chunk" from the left side of R. Similarly, define r_2, r_3, r_4, which represent chunks taken from the right, bottom, and top of R, respectively.

Now consider the risk of R_S, which is the probability that the hypothesis R_S makes an incorrect prediction:

$$\mathcal{P}(R_S) = \mathcal{P}(1_{R_S}(X) \neq Y \mid X \in R \setminus R_S) = \mathcal{P}(X \in R \setminus R_S).$$

If $\mathcal{P}(R_S) \geq \epsilon$, then $R_S \cap r_i = \emptyset$ for some $i \in \{1, 2, 3, 4\}$ because, otherwise, the set $R \setminus R_S$ would be covered by the union of the four chunks $r_1, \ldots, r_4$, which would contradict the assumption that the error is larger than ϵ.

Thus, we have the following bound on the probability that $\mathcal{P}(R_S) \geq \epsilon$:

$$\mathcal{P}(R_S \geq \epsilon) \leq \mathcal{P}\left(\bigcup_{i=1}^{4} \{R_S \cap r_i = \emptyset\}\right).$$

By the union bound, this is less than or equal to:

$$\sum_{i=1}^{4} \mathcal{P}(R_S \cap r_i = \emptyset).$$

Next, we can further bound this by the probability that all points in the sample S fall outside of r_i:

$$\sum_{i=1}^{4} \mathcal{P}(X_1 \in r_i, Y_1 \neq 1_{r_i}(X_1), \ldots, X_m \in r_i, Y_m \neq 1_{r_i}(X_m)).$$

Since the points $(X_1, Y_1), \ldots, (X_m, Y_m)$ are i.i.d., we can simplify this as:

$$\sum_{i=1}^{4} \mathcal{P}(X \in r_i, Y \neq 1_{r_i}(X))^m.$$

Now, consider the probability $\mathcal{P}(X \in r_i, Y \neq 1_{r_i}(X))$. Since $Y = 1$ with probability $1 - \eta$ for $X \in r_i$, we have:

$$\mathcal{P}(X \in r_i, Y \neq 1_{r_i}(X)) = 1 - \mathcal{P}(X \in r_i, Y = 1).$$

This equals:

$$1 - (1 - \eta)\mathcal{P}(X \in r_i) \geq (1 - \eta)\epsilon/4.$$

Hence, the probability of error on r_i is bounded by:

$$\mathcal{P}(X \in r_i, Y \neq 1_{r_i}(X))^m \leq \exp(-m(1 - \eta)\epsilon/4).$$

Summing over all four chunks r_1, r_2, r_3, r_4, we obtain the following bound:

$$\sum_{i=1}^{4} \mathcal{P}(X \in r_i, Y \neq 1_{r_i}(X))^m \leq 4 \exp(-m(1 - \eta)\epsilon/4).$$

Thus, we have:

$$\mathcal{P}(R_S \geq \epsilon) \leq 4 \exp(-m(1 - \eta)\epsilon/4).$$

Finally, to ensure that the probability of error is at most δ, we set the right-hand side smaller than δ, leading to the sample complexity bound:

$$m \geq \frac{4}{(1 - \eta)\epsilon} \log\left(\frac{4}{\delta}\right).$$

This shows that the sample complexity required to PAC learn in this noisy scenario, with noise rate η, is $O\left(\frac{1}{(1-\eta)\epsilon} \log\left(\frac{1}{\delta}\right)\right)$. Hence, we can still achieve agnostic PAC learning despite the label noise, provided we have enough samples. ■

Solution 6:

Observe that the subset $\{1\} \subseteq X$ is trivially shattered by $\{c_0, c_1\}$. Additionally, the subset $\{12, 23\} \subseteq X$ is shattered by $\{c_0, c_1, c_2, c_3\}$, as each number has a distinct combination of digits.

Now, consider a general subset $\{x, y, z\} \subseteq X$. Suppose $\{x, y, z\}$ is shattered by C. This would imply the existence of indices $i, j, k, l \in \{0, 1, 2, \ldots, 9\}$ such that:

1. $x \in c_i$, but $y, z \notin c_i$,
2. $x, y \in c_j$, but $z \notin c_j$,
3. $x, z \in c_k$, but $y \notin c_k$,
4. $x, y, z \in c_l$.

From condition 1, we deduce that x must have a digit that is not in y or z. Let this digit be d_{xy}. Similarly, condition 2 implies that x and y share a digit not in z, say d_{xz}, and condition 3 implies that x and z share a digit not in y, say d_{yz}. Lastly, condition 4 implies that x, y, z all share a common digit, say d_{xyz}.

However, observe that none of $d_{xy}, d_{xz}, d_{yz}, d_{xyz}$ can be equal, as each must represent a distinct condition on the digits of x, y, z. But since each number in X has at most three digits, this leads to a contradiction (there are only three digits available, but we require four distinct digits to satisfy all the conditions).

Therefore, no subset of size 3 can be shattered by C. Since subsets of size 2 can be shattered, we conclude that the VC dimension of C is 2. ■

Solution 7:

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables taking values in $[0,1]$ with mean μ. Define $\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$. We aim to show that for any ϵ such that $\mu + \epsilon < 1$, the following inequality holds:

$$\mathbb{P}(\overline{X} \geq \mu + \epsilon) \leq e^{-mKL^+(\mu+\epsilon\|\mu)},$$

where $KL^+(p\|q)$ is the KL divergence between two Bernoulli distributions, given by

$$KL^+(p\|q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}.$$

To derive the desired bound, we apply Chernoff's method. Consider the probability we seek to bound:

$$\mathbb{P}\left(\overline{X} \geq \mu + \epsilon\right) = \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} X_i \geq \mu + \epsilon\right) = \mathbb{P}\left(\sum_{i=1}^{m} X_i \geq m(\mu+\epsilon)\right).$$

Using the exponential Markov inequality, for any $t > 0$, we have:

$$\mathbb{P}\left(\sum_{i=1}^{m} X_i \geq m(\mu+\epsilon)\right) \leq \mathbb{E}\left[e^{t\sum_{i=1}^{m} X_i}\right] e^{-tm(\mu+\epsilon)}.$$

Since $X_1, X_2, \ldots, X_m$ are independent and identically distributed, the moment generating function of $\sum_{i=1}^{m} X_i$ factorizes as:

$$\mathbb{E}\left[e^{t\sum_{i=1}^{m} X_i}\right] = \prod_{i=1}^{m}\mathbb{E}\left[e^{tX_i}\right] = \left(\mathbb{E}\left[e^{tX_1}\right]\right)^m.$$

Thus, the bound becomes:

$$\mathbb{P}(\overline{X} \geq \mu + \epsilon) \leq \left(\mathbb{E}\left[e^{tX_1}\right]\right)^m e^{-tm(\mu+\epsilon)}.$$

Next, to minimize this bound, we optimize the exponent by choosing the value of t that minimizes the function $\mathbb{E}\left[e^{tX_1}\right] - t(\mu+\epsilon)$. Let $\psi(t) = \ln\mathbb{E}\left[e^{tX_1}\right]$. We now aim to minimize the exponent:

$$\psi(t) - t(\mu+\epsilon).$$

To find the optimal t, we differentiate with respect to t and set the derivative to zero:

$$\frac{d}{dt}\left(\psi(t) - t(\mu+\epsilon)\right) = 0 \quad \Rightarrow \quad \psi'(t) = \mu + \epsilon.$$

This equation characterizes the optimal choice of t.

For random variables $X_i \in [0,1]$, the cumulant generating function $\psi(t)$ is related to the Kullback-Leibler divergence between two Bernoulli distributions. Specifically, the minimum value of $\psi(t) - t(\mu+\epsilon)$ corresponds to $-KL^+(\mu+\epsilon\|\mu)$, where:

$$KL^+(\mu+\epsilon\|\mu) = (\mu+\epsilon)\ln\frac{\mu+\epsilon}{\mu} + (1-\mu-\epsilon)\ln\frac{1-\mu-\epsilon}{1-\mu}.$$

Thus, the probability bound becomes:

$$\mathbb{P}(\overline{X} \geq \mu + \epsilon) \leq e^{-mKL^+(\mu+\epsilon\|\mu)}.$$

Finally, we deduce Hoeffding's inequality from this result. Hoeffding's inequality provides a bound for sums of bounded independent random variables. Specifically, for random variables $X_i \in [0, 1]$, Hoeffding's inequality states:

$$\mathbb{P}(\overline{X} \geq \mu + \epsilon) \leq e^{-2m\epsilon^2}.$$

To see how this follows from the KL-divergence bound, we approximate $KL^+(\mu + \epsilon \| \mu)$ for small ϵ. Using a second-order Taylor expansion around $\epsilon = 0$, we find:

$$KL^+(\mu + \epsilon \| \mu) \approx \frac{\epsilon^2}{2\mu(1-\mu)}.$$

Since $\mu(1-\mu) \leq \frac{1}{4}$ for $\mu \in [0, 1]$, we obtain the inequality:

$$KL^+(\mu + \epsilon \| \mu) \geq 2\epsilon^2.$$

Therefore, the bound:

$$\mathbb{P}(\overline{X} \geq \mu + \epsilon) \leq e^{-mKL^+(\mu+\epsilon\|\mu)}$$

implies Hoeffding's inequality:

$$\mathbb{P}(\overline{X} \geq \mu + \epsilon) \leq e^{-2m\epsilon^2}.$$

■

Solution 8:

We need to show that when q is fixed, $KL^+(p\|q)$ is a convex function of p, and when p is fixed, $KL^+(p\|q)$ is a convex function of q.

First, consider q fixed. Let

$$f(p) = KL^+(p\|q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}.$$

To establish convexity, we need to compute the second derivative of $f(p)$ with respect to p. The first derivative of $f(p)$ is:

$$f'(p) = \frac{d}{dp}\left(p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}\right)$$

Applying the product and chain rules, we obtain:

$$f'(p) = \ln\frac{p}{q} + 1 - \ln\frac{1-p}{1-q} - 1 = \ln\frac{p}{q} - \ln\frac{1-p}{1-q}.$$

Next, we compute the second derivative:

$$f''(p) = \frac{d}{dp}\left(\ln\frac{p}{q} - \ln\frac{1-p}{1-q}\right)$$

This gives:

$$f''(p) = \frac{1}{p} + \frac{1}{1-p}.$$

Since $\frac{1}{p} + \frac{1}{1-p} > 0$ for all $p \in (0,1)$, the second derivative is strictly positive, implying that $f(p)$ is convex for all $p \in (0,1)$. Therefore, $KL^+(p\|q)$ is a convex function of p when q is fixed.

Now, consider p fixed. Let

$$h(q) = KL^+(p\|q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}.$$

We now compute the second derivative of $h(q)$ with respect to q to show convexity. The first derivative of $h(q)$ is:

$$h'(q) = \frac{d}{dq}\left(p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}\right)$$

Using the product and chain rules, we obtain:

$$h'(q) = -\frac{p}{q} + \frac{1-p}{1-q}.$$

Next, we compute the second derivative:

$$h''(q) = \frac{d}{dq}\left(-\frac{p}{q} + \frac{1-p}{1-q}\right)$$

This gives:

$$h''(q) = \frac{p}{q^2} + \frac{1-p}{(1-q)^2}.$$

Since $\frac{p}{q^2} + \frac{1-p}{(1-q)^2} > 0$ for all $p, q \in (0,1)$, the second derivative is strictly positive, which implies that $h(q)$ is convex for all $q \in (0,1)$. Therefore, $KL^+(p\|q)$ is a convex function of q when p is fixed.

Thus, we have shown that $KL^+(p\|q)$ is convex in p when q is fixed and convex in q when p is fixed. ■

Solution 9:

Let $p(h)$ and $q(h)$ be two functions defined on the hypothesis space $\mathcal{H}$, where $p(h) \in (0,1)$ and $q(h) \in (0,1)$, and let Q be a probability distribution on $\mathcal{H}$. We want to show that

$$KL^+\left(\mathbb{E}_{h\sim Q}[p(h)] \| \mathbb{E}_{h\sim Q}[q(h)]\right) \leq \mathbb{E}_{h\sim Q}\left[KL^+(p(h)\|q(h))\right].$$

From problem 8, we have, the function $KL^+(p\|q)$ is a convex function of p when q is fixed and a convex function of q when p is fixed. We shall use Jensen's inequality, which states that for any convex function $f(X)$ and a random variable X with distribution Q, the following inequality holds:

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

Applying this to the convex function $f(p(h), q(h)) = KL^+(p(h)\|q(h))$ and the distribution Q, we get

$$KL^+\left(\mathbb{E}_{h\sim Q}[p(h)] \| \mathbb{E}_{h\sim Q}[q(h)]\right) = f\left(\mathbb{E}_{h\sim Q}[p(h)], \mathbb{E}_{h\sim Q}[q(h)]\right) \leq \mathbb{E}_{h\sim Q}[f(p(h), q(h))].$$

Since $f(p(h), q(h)) = KL^+(p(h)\|q(h))$, this simplifies to

$$KL^+\left(\mathbb{E}_{h\sim Q}[p(h)] \| \mathbb{E}_{h\sim Q}[q(h)]\right) \leq \mathbb{E}_{h\sim Q}\left[KL^+(p(h)\|q(h))\right],$$

which completes the solution. ■