18.657 PS 1 SOLUTIONS

1. Problem 1

(1) We expand in terms of conditional probabilities:

$$\begin{aligned}\mathbb{P}(X \in U) &= \mathbb{P}(X \in U \mid Y = 0)\,\mathbb{P}(Y = 0) + \mathbb{P}(X \in U \mid Y = 1)\,\mathbb{P}(Y = 1)\\ &= \pi_0\,\mathbb{P}(X \in U \mid Y = 0) + \pi_1\,\mathbb{P}(X \in U \mid Y = 1).\end{aligned}$$

Passing to densities:

$$\begin{aligned}p_X &= \pi_0\, p_{X\mid Y=0} + \pi_1\, p_{X\mid Y=1}\\ &= \pi_0(2\pi)^{-d/2}(\det\Sigma_0)^{-1/2}\exp\left(-\frac{1}{2}(x-\mu_0)^\top\Sigma_0^{-1}(x-\mu_0)\right)\\ &\quad+\pi_1(2\pi)^{-d/2}(\det\Sigma_1)^{-1/2}\exp\left(-\frac{1}{2}(x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1)\right).\end{aligned}$$

(2) By Bayes' rule, we have $\mathbb{P}(Y = y \mid X = x) \propto p_{X\mid Y=y}(x)\,\mathbb{P}(Y = y)$. Thus, given $X = x$, the value of the Bayes classifier is 0 if

$$\pi_0 p_{X\mid Y=0} > \pi_1 p_{X\mid Y=1},$$

and is otherwise 1. Cancelling $(2\pi)^{-d/2}(\det\Sigma)^{-1/2}$ on both sides, and taking logs, the inequality reads

$$\log\pi_0 - \frac{1}{2}(x-\mu_0)^\top\Sigma^{-1}(x-\mu_0) > \log\pi_1 - \frac{1}{2}(x-\mu_1)^\top\Sigma^{-1}(x-\mu_1).$$

This asks whether x is at least $\log\pi_0 - \log\pi_1$ units closer to μ_0 than μ_1, in distance measured by the quadratic form $\frac{1}{2}\Sigma^{-1}$. By the PSD assumption, this distance matches Euclidean distance after a linear transformation (given by the Cholesky factorization of Σ^{-1}). Then, geometrically, the boundary forms a hyperplane, and the two regions $\{h^* = 0\}$ and $\{h^* = 1\}$ form half-spaces.

(3) The inequality defining the Bayes classifier remains quadratic in x, so the hypersurface separating $\{h^* = 0\}$ from $\{h^* = 1\}$ is a quadric. For example, with

$$\Sigma_0 = \begin{pmatrix}1 & 0\\ 0 & 2\end{pmatrix},\quad \Sigma_1 = \begin{pmatrix}2 & 0\\ 0 & 1\end{pmatrix},\quad \pi_0 \neq \pi_1,$$

the separating curve is a hyperbola.

2. Problem 2

(1) We see that $\mathcal{C}$ can shatter $2d+1$ points on a circle: proceeding clockwise around the circle, we pass a line from last of each run of consecutive 'included' points to start of the next such run, for a total of at most d lines, and define our polygon by these lines (which extend beyond their defining points).

Suppose we have $2d+2$ points. If any point is in the convex hull of the others, then $\mathcal{C}$ fails to shatter: any set in $\mathcal{C}$ is convex, and thus can not include the extreme points while excluding an interior point. Otherwise, number the points in their clockwise order as vertices of the convex hull, and consider an alternating sign pattern. In order to shatter, there must be a face of the polygon from $\mathcal{C}$ passing between each excluded point and its two included neighbors; these $d+1$ faces must all be distinct, or else we could show that the given points violated our convexity assumption; but no convex polygon with d vertices has more than d faces, a contradiction. So $\mathcal{C}$ can not shatter $2d+2$ points.

(2) The convex compact sets include in particular the convex polygons of the previous section, which we see can shatter any number of points (once we allow arbitrarily many polygon vertices). Thus the VC dimension is infinite.

(3) We know that $\mathcal{C}$ shatters some $n = \mathsf{VC}(\mathcal{C})$ points. Then some set in $\mathcal{C}$ achieves each of the 2^n inclusion patterns. In particular there exist at least 2^n sets in $\mathcal{C}$. So $n \le \log_2(\operatorname{card}\mathcal{C})$.

(4) On the space $\mathbb{R}$, the class of intervals $(-\infty, t]$ for $t \in \mathbb{R}$ is infinite. This class certainly shatters the point 0 (the intervals $(-\infty, -1]$ and $(-\infty, 0]$ suffice), while it fails to shatter any two points $a < b$, since no such interval contains b but not a. Thus the VC dimension is 1.

3. Problem 3

(1) We begin by computing mean and variance:

$$\begin{aligned}\mathbb{E}[\hat{F}_n(t)] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(X_i \le t) = \frac{1}{n}\sum_{i=1}^{n}F(t) = F(t),\\ \mathrm{Var}[\hat{F}_n(t)] &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[\mathbb{1}(X_i \le t)]\\ &= \frac{1}{n}\mathrm{Var}[\mathbb{1}(X_1 \le t)]\\ &= \frac{1}{n}(\mathbb{E}[\mathbb{1}(X_1 \le t)^2] - \mathbb{E}[\mathbb{1}(X_1 \le t)]^2) = \frac{1}{n}(F(t) - F(t)^2).\end{aligned}$$

Let $\varepsilon > 0$ be given. As the $\hat{F}_n(t)$ are averages of independent 0–1 random variables, we can apply Hoeffding's inequality:

$$\begin{aligned}\mathbb{P}(|\hat{F}_n(t) - F(t)| \ge \varepsilon) &\le 2\exp(-2n\varepsilon^2),\\ \sum_{n=1}^{\infty}\mathbb{P}(|\hat{F}_n(t) - F(t)| \ge \varepsilon) &\le \sum_{n=1}^{\infty}2\exp(-2n\varepsilon) = \frac{2}{1-\exp(-2\varepsilon)} < \infty,\end{aligned}$$

so that by Borel–Cantelli, almost surely only finitely many of the events $|\hat{F}_n(t) - F(t)| \ge \varepsilon$ occur. Thus $F_n(t) \to F(t)$ almost surely.

(Alternative: $\hat{F}_n(t)$ is an average of n iid samples of $\mathrm{Bern}(F(t))$; the result follows from the strong law of large numbers. Of course, the proof of that law uses Borel–Cantelli.)

(2) We can view this is a question of empirical measure versus true measure on the class $\mathcal{C}$ of half-line intervals $(\infty, t]$, which have VC dimension 1 (see the solution to 2(d)). The VC inequality now asserts that

$$\begin{aligned}\sup_{t\in\mathbb{R}}|\hat{F}_n(t) - F(t)| &= \sup_{A\in\mathcal{C}}|\mu_n(A) - \mu(A)|\\ &\le 2\sqrt{\frac{2\log(2en)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}\\ &\le \left(2\sqrt{\frac{2\log 4e}{\log 2}} + \frac{1}{\sqrt{2}}\right)\sqrt{\frac{\log(n/\delta)}{n}}\end{aligned}$$

for all $n \ge 2$.

4. Problem 4

(1) Changing the value X_i can only change B_n by at most 1: putting the ith item in a new bin is always an option. Applying the bounded differences inequality yields exactly the result:

$$\mathbb{P}(|B_n - \mathbb{E}B_n| > t) \le 2\exp(-2t^2/n).$$

(2) (a) By the triangle inequality, changing X_i will change $\|\bar{X}\|$ by at most $2/n$. By the (one-sided) bounded differences inequality,

$$\mathbb{P}(\|\bar{X}\| - \mathbb{E}\|\bar{X}\| \ge t) \le \exp\left(\frac{-2t^2}{n(2/n)^2}\right) = \exp(-nt^2/2).$$

(b) We apply Jensen's inequality:

$$\begin{aligned}\mathbb{E}[\|\bar{X}\|] &\le \mathbb{E}[\|\bar{X}\|^2]^{1/2} \\ &= \left(\sum_{i=1}^{d} \mathbb{E}[\bar{X}(i)^2]\right)^{1/2}.\end{aligned}$$

By the assumption $\mathbb{E}[X] = 0$, the cross terms of the squared average will cancel in expectation:

$$\mathbb{E}[\bar{X}(i)^2] = \frac{1}{n^2}\sum_{j=1}^{n} \mathbb{E}[X_j(i)^2] = \frac{1}{n}\mathbb{E}[X(i)^2].$$

Returning to our computation:

$$\begin{aligned}\mathbb{E}[\|\bar{X}\|] &\le \left(\sum_{i=1}^{d} \frac{1}{n}\mathbb{E}[X(i)^2]\right)^{1/2} \\ &= \left(\frac{1}{n}\mathbb{E}[\|X\|^2]\right)^{1/2} \\ &\le \frac{1}{\sqrt{n}}.\end{aligned}$$

(c) Combining (a) and (b), we have

$$\begin{aligned}\mathbb{P}(\|\bar{X}\| \ge t) &\le \exp(-n(t - 1/\sqrt{n})^2/2) \\ &= \exp\left(-\frac{1}{2}(t\sqrt{n})^2 + t\sqrt{n} - \frac{1}{2}\right).\end{aligned}$$

The polynomial $-\frac{1}{2}x^2 + x - \frac{1}{2}$ is bounded above by $\frac{1}{2} - x^2/4$. Thus,

$$\mathbb{P}(\|\bar{X}\| \ge t) \le e^{1/2}\exp(-nt^2/4) \le 2\exp(-nt^2/4).$$

(3) Let $\sigma_1, \ldots, \sigma_n$ be independent symmetric Rademacher random variables. By symmetry of X_i, the distribution of $\bar{X}$ is the same as that of $\langle \sigma, X\rangle/n$. Conditioned on X, Hoeffding's inequality yields

$$\mathbb{P}_\sigma\left(\frac{\langle \sigma, X\rangle}{n\sqrt{V}} > t\right) \le \exp(-nt^2/2).$$

By the law of total probability,

$$\begin{aligned}\mathbb{P}_X\left(\frac{\bar{X}}{\sqrt{V}} > t\right) = \mathbb{P}_{X,\sigma}\left(\frac{\langle \sigma, X\rangle}{n\sqrt{V}} > t\right) &= \mathbb{E}_X\,\mathbb{P}_\sigma\left(\frac{\langle \sigma, X\rangle}{n\sqrt{V}} > t\right) \\ &\le \mathbb{E}_X \exp(-nt^2/2) \\ &= \exp(-nt^2/2).\end{aligned}$$

18.657 Mathematics of Machine Learning
Fall 2015