Mathematics of Machine Learning
Homework 7 - Solutions

Instructors: Afonso S. Bandeira & Nikita Zhivotovskiy
Course Coordinator: Pedro Abdalla Teixeira

April 23, 2021

Try to solve the questions before looking to the answers. Every item must be proved rigorously. Starred problems are harder.

Problem 1

Let X be a random variable. Prove the following classical results

(a) Markov's inequality: For every $a > 0$, prove that

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}|X|}{a}$$

(b) Chebyshev's inequality: For every $a > 0$ and $p \geq 1$, prove that

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}|X|^p}{a^p}.$$

(Hint: Use that $\{a \in \mathbb{R} : |X| \geq a\} = \{a \in \mathbb{R} : |X|^p \geq a^p\}$.)

(c) Integral identity/Layer cake representation: For a nonnegative integrable random variable Y, prove that

$$\mathbb{E}Y = \int_0^\infty \mathbb{P}(Y \geq t)dt.$$

(d) Generalize the result above for an integrable random variable X that may take negative values. (Hint: Split the random variable in a sum of two nonnegative random variables)

Solution 1

For the rest of this exercise, the indicator function of a event A is denoted by $\mathbb{1}_A$.

(a) We split $|X|$ into the sum $|X|\mathbb{1}_{|X|<a} + |X|\mathbb{1}_{|X|\geq a}$. Clearly, it follows that $|X| \geq |X|\mathbb{1}_{|X|\geq a}$. This implies that

$$\mathbb{E}|X| \geq a\mathbb{E}[\mathbb{1}_{|X|\geq a}] = a\mathbb{P}(|X| \geq a).$$

(b) Observe that, for all $p \geq 1$, $\mathbb{P}(|X| \geq a) = \mathbb{P}(|X|^p \geq a^p)$, Chebyshev inequality follows by applying Markov inequality (letter "a").

(c) Let's denote μ by the law of Y. We write

$$\mathbb{E}Y = \int_0^\infty y d\mu = \int_0^\infty \int_0^\infty \mathbb{1}_{t\leq y} dt d\mu = \int_0^\infty \mathbb{P}(Y \geq t)dt.$$

In the last step we used Fubini-Tonelli theorem to switch the integrals.

(d) We split X into $X\mathbb{1}_{X\leq 0} + X\mathbb{1}_{X\geq 0}$. We write

$$\mathbb{E}X\mathbb{1}_{X<0} = \mathbb{E}X\mathbb{1}_{-X>0} = \int_0^\infty \mathbb{P}(-X \geq t)dt = -\int_{-\infty}^0 \mathbb{P}(X \geq t)dt.$$

The second equality follows from letter "c" and the last equality follows by simple change of variables. By applying letter "c" again to the nonnegative random variable $X\mathbb{1}_{X\geq 0}$ we get

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X \geq t)dt - \int_{-\infty}^0 \mathbb{P}(X \geq t)dt.$$

Problem 2

In this exercise we denote $Var(X)$ by the variance of the random variable X. Consider the random variables $X_1, \dots, X_n$ and the random sum $Z := \sum_{i=1}^n X_i$.

(a) If $X_1, \dots, X_n$ are independent, prove that $Var(Z) = \sum_{i=1}^n Var(X_i)$.

(b) Is independence a necessary condition for the formula above?

(c) Let $\mathbb{E}_i$ denote the conditional expectation operator, conditioned on $X_1, \dots, X_i$ with the convention that $\mathbb{E}_0$ is the expectation with respect to all $X_1, \dots, X_n$. Check that

$$Z - \mathbb{E}Z = \sum_{i=1}^n \Delta_i,$$

where $\Delta_i := \mathbb{E}_i Z - \mathbb{E}_{i-1} Z$.

(d) Prove that $Var(Z) = \sum_{i=1}^n \mathbb{E}\Delta_i^2$

Solution 2

(a) Recall that if two random variables X and Y are independent, then $\mathbb{E}[XY] = (\mathbb{E}X)(\mathbb{E}Y)$. In particular, if either X or Y is mean zero, then $\mathbb{E}[XY] = 0$. By definition of variance, we may write the variance of Z as

$$\sum_{i,j=1}^{n} \mathbb{E}[(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)] = \sum_{i=1}^{n} \mathbb{E}[(X_i - \mathbb{E}X_i)^2] = \sum_{i=1}^{n} Var(X_i).$$

In the last step we used the fact that $X_i - \mathbb{E}X_i$ and $X_j - \mathbb{E}X_j$ are independent mean zero random variables.

(b) The answer is no. Observe that in the proof above we used the fact that the random variables are uncorrelated, this is a weaker requirement than independence. For example, take X uniformly distributed in the interval $[-1.1]$ and take $Y = X^2$. Their correlation is zero but they are not independent.

(c) The telescopic sum $\sum_{i=1}^{n} \Delta_i = \mathbb{E}_n Z - \mathbb{E}_0 Z$. By construction, $\mathbb{E}_0 Z = \mathbb{E}Z$ and conditioning on $X_1, \dots, X_n$, Z is non-random, so $\mathbb{E}_n Z = Z$.

(d) Observe that

$$Var(Z) = \mathbb{E}[(\sum_{i=1}^{n} \Delta_i)^2] = \sum_{i=1}^{n} \mathbb{E}\Delta_i^2 + 2 \sum_{j>i} \mathbb{E}[\Delta_i \Delta_j].$$

It remains to prove that the second term (RHS) is zero. Indeed, $\mathbb{E}_i \Delta_i \Delta_j = \Delta_i \mathbb{E}\Delta_j = 0$, therefore $\mathbb{E}[\Delta_i \Delta_j] = 0$ by iterated law of expectation.

Problem 3

Consider a binary classification problem where $\mathcal{X} = \mathbb{N}$ and the class $\mathcal{F}$ consists of all classifiers that are equal to one on exactly one integer.

(a) For the class $\mathcal{F}$, construct a sample compression scheme of size one.

(b) Using a direct computation, prove that $\mathcal{F}$ is PAC learnable with the sample complexity $n(\varepsilon, \delta) = \lceil \frac{1}{\varepsilon} \log \frac{1}{\delta} \rceil$.

(Hint: It might be convenient to use a reconstruction function ρ that is not restricted to $\mathcal{F}$.)

Solution 3

(a) Let $f^* \in \mathcal{F}$ denote a unknown classifier that we want to learn. Notice that it can be determined if we collect a sample of the type $(X, 1)$, i.e, a sample with label 1. This holds because all classifiers in the class are singletons. Our compression function κ maps the sample $S_n := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ into a single point as follows: If there exists $j \in [n]$ with $Y_j = 1$, then the compression function maps $\kappa(S_n) = X_j$, otherwise it maps the sample to zero. The size of the compression scheme is one because we only save one point. The reconstruction function ρ maps 0 into all zero classifier (that does not belong to the class $\mathcal{F}$) and maps X_j into the classifier $\hat{f}_j$ defined as

$$\hat{f}_j(X) = \begin{cases} 1, X = X_j \\ 0, \text{otherwise} \end{cases}$$

Clearly $\hat{f} = \rho(\kappa(S_n))$ satisfies $\hat{f}(X_i) = f^*(X_i)$.

(b) Recall that the risk is $R(\hat{f}) = \mathbb{P}(\hat{f}(X) \neq f^*(X))$. Using our sample compression scheme developed in item "a", the only error that might occur is when $f^*(X) = 1$ and, for all X_i, $\hat{f}(X_i) = 0$. So the risk is only nonzero if $\hat{f}$ is all zero classifier and in such case $R(\hat{f}) = \mathbb{P}(\hat{f}(x) \neq f^*(x)) = \mathbb{P}(X \text{ occurs})$. We assume that the latter probability is at least ε, otherwise the risk is never larger than ε. Now we can write,

$$\mathbb{P}(R(\hat{f}) > \varepsilon) \leq \mathbb{P}\left(\bigcap_{i=1}^{n}(X_i \neq X)\right) \leq \prod_{i=1}^{n}\mathbb{P}(X_i \neq X) \leq \prod_{i=1}^{n}(1-\varepsilon).$$

We apply numeric inequality $1 - x \leq e^{-x}$ together with the fact that $n = \lceil \frac{1}{\varepsilon}\log\frac{1}{\delta} \rceil$ to conclude that

$$\mathbb{P}(R(\hat{f}) > \varepsilon) \leq e^{-\varepsilon n} \leq \delta.$$

We conclude that the class is PAC learnable with sample complexity $n = \lceil \frac{1}{\varepsilon}\log\frac{1}{\delta} \rceil$.