CSC 665: Homework 1

Chicheng Zhang

October 12, 2019

Please complete the following set of exercises **on your own**. The homework is due **on Oct 1, 12:30pm, on Gradescope**. You are free to cite existing theorems from the textbook and course notes.

Problem 1

For a random variable Z with mean $\mathbb{E}Z = 0$, we call Z is v-subgaussian, if

$$\psi_Z(t) = \ln \mathbb{E}e^{tZ} \leq \frac{vt^2}{2}.$$

Show the following:

1. If Z has Gaussian distribution $\mathrm{N}(0, \sigma^2)$, then Z is σ^2-subgausssian.
2. If Z take values within interval $[a, b]$, then Z is $\frac{(b-a)^2}{4}$-subgaussian.
3. If $Z_1, \ldots, Z_n$ are independent, and each Z_i is v_i subgaussian, then $\sum_{i=1}^n Z_i$ is $\sum_{i=1}^n v_i$-subgaussian.
4. If Z is v-subgaussian, then
$$\mathbb{P}(|Z| \geq t) \leq 2\exp\left\{-\frac{t^2}{2v}\right\}.$$

Problem 2

In this exercise we give an alternative proof of the Chernoff bound for Bernoulli random variables: suppose $X_1, \ldots, X_n$ are iid and from Bernoulli(p), define $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, then,

$$\mathbb{P}(\bar{X} \geq q) \leq \exp\{-n\,\mathrm{kl}(q, p)\}, q \geq p, \tag{1}$$

$$\mathbb{P}(\bar{X} \leq q) \leq \exp\{-n\,\mathrm{kl}(q, p)\}, q \leq p. \tag{2}$$

1. Show that
$$\mathbb{P}(\bar{X} \geq q) = \sum_{m:m\geq nq} \binom{n}{m} p^m (1-p)^{n-m}.$$
2. Use the elementary inequality that $\binom{n}{m} q^m (1-q)^{n-m} \leq 1$, show that for $m \geq nq$,
$$\binom{n}{m} p^m (1-p)^{n-m} \leq \left(\frac{p}{q}\right)^{nq} \left(\frac{1-p}{1-q}\right)^{n(1-q)}.$$

3. Use the above two items to conclude that $\mathbb{P}(\bar{X} \geq q) \leq (n+1)\exp\bigl\{-n\,\mathrm{kl}(q,p)\bigr\}$.

4. Note that compared to Equation 1, the above bound is has an additional factor of n on the right hand side. Use the elementary inequality $\sum_{m \geq nq} \binom{n}{m} q^m (1-q)^{n-m} \leq 1$ as a starting point, along with insights you gained from items 1 and 2 to show Equation (1).

5. Repeat the proof for the lower tail bound (Equation (2)).

Problem 3

In this exercise we will use basic concentration inequalities to show that, we can find exponentially many points on the unit sphere in $\mathbb{R}^d$ that are far away from each other. Specifically, consider n random vectors $X_1, X_2, \ldots, X_n$ in $\mathbb{R}^d$, where for each i, $X_i = \frac{1}{\sqrt{d}}(Z_{i,1}, \ldots, Z_{i,d})$. Here $\bigl\{Z_{i,j}\bigr\}_{i \in \{1,\ldots,n\}, j \in \{1,\ldots,d\}}$'s are all independent and identically distributed, and $Z_{i,j}$ takes value 1 with probability 1/2, and takes value -1 with probabilty 1/2.

1. Check that all X_i's has unit length, i.e. $\|X_i\|_2 = 1$.

2. Use Hoeffding's Inequality to show that for any fixed pair $i, j \in \{1, \ldots, n\}$, $i \neq j$,

$$\mathbb{P}(|\left\langle X_i, X_j \right\rangle| \geq \frac{1}{2}) \leq 2\exp\left\{-\frac{d}{8}\right\}.$$

3. Suppose $n = \exp\left\{\frac{d}{32}\right\}$. Show that with nonzero probability, for all pairs $i, j \in \{1, \ldots, n\}$, $i \neq j$, the angle between X_i and X_j is in $[\frac{\pi}{3}, \frac{2\pi}{3}]$.

Problem 4

Suppose D is a distribution over $[0,1] \times \{-1,+1\}$ such that D_X, the marginal of D over $\mathcal{X} = [0,1]$, is uniform. In addition,

$$P(Y = +1|x) = \begin{cases} 0 & x \leq \frac{1}{2}, \\ 1 & x > \frac{1}{2} \end{cases},$$

i.e. the distribution is separable by a threshold classifier with threshold $\frac{1}{2}$. Suppose training examples $(X_1, Y_1), \ldots, (X_n, Y_n)$ are drawn iid from D. Now consider the following classifier $\hat{h}$:

$$\hat{h}(x) = \begin{cases} Y_i & x = X_i \text{ for some } i \in \{1, \ldots, n\}, \\ -1 & \text{otherwise.} \end{cases}$$

(For simplicity, assume that all X_i's are distinct, which also happens with probability 1.)

1. Calculate $\mathrm{err}(\hat{h}, S)$.

2. Calculate $\mathrm{err}(\hat{h}, D)$. What is the value of $\mathrm{err}(\hat{h}, S) - \mathrm{err}(\hat{h}, D)$?

3. It may be tempting to use following argument to argue the concentration of $\mathrm{err}(\hat{h}, S)$ to $\mathrm{err}(\hat{h}, D)$. Define random variables $Z_i = \mathbf{1}(\hat{h}(X_i) \neq Y_i)$ for all i in $\{1, \ldots, n\}$, therefore, Hoeffding's inequality, with probability $1 - \delta$,

$$|\,\mathrm{err}(\hat{h}, S) - \mathrm{err}(\hat{h}, D)| \leq \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}.$$

Does this contradict the results we got from item 2? Why?

Problem 5

In this exercise, we will unify the analysis of $O(\frac{1}{\epsilon})$-style sample complexity for the realizable case and the $O(\frac{1}{\epsilon^2})$-style sample complexity for the agnostic case, by revisiting the empirical risk minimization algorithm. Suppose $\mathcal{H}$ is a finite hypothesis class, D is a distribution over labeled examples, and S is a training set of size m drawn iid from D. Denote by $\nu^\star = \min_{h \in \mathcal{H}} \mathrm{err}(h, D)$ as the optimal generalization error, and $\hat{h}$ the output of the empirical risk minimzation algorithm.

1. Use Chernoff bound for Bernoulli random variables, show that for a fixed classifier h, with probability $1 - \delta$,
$$\mathrm{kl}(\mathrm{err}(h, S), \mathrm{err}(h, D)) \leq \frac{\ln \frac{2}{\delta}}{m}.$$

2. Use the above reasoning to conclude that with probability $1 - \delta$, for all classifiers h in $\mathcal{H}$,
$$|\,\mathrm{err}(h, S) - \mathrm{err}(h, D)| \leq \sqrt{2 \max(\mathrm{err}(h, S), \mathrm{err}(h, D)) \frac{\ln \frac{2|\mathcal{H}|}{\delta}}{m}}.$$
(Hint: you can use the fact that $\mathrm{kl}(q, p) \geq \frac{(q-p)^2}{2\max(p,q)}$.)

3. Show that with probability $1 - \delta$, for all classifiers h in $\mathcal{H}$,
$$\mathrm{err}(h, S) \leq \mathrm{err}(h, D) + \sqrt{\mathrm{err}(h, D) \frac{2 \ln \frac{2|\mathcal{H}|}{\delta}}{m}} + \frac{2 \ln \frac{2|\mathcal{H}|}{\delta}}{m},$$
$$\mathrm{err}(h, D) \leq \mathrm{err}(h, S) + \sqrt{\mathrm{err}(h, S) \frac{2 \ln \frac{2|\mathcal{H}|}{\delta}}{m}} + \frac{2 \ln \frac{2|\mathcal{H}|}{\delta}}{m}.$$
(Hint: you can use the elementary fact that for $A, B, C > 0$, $A \leq B + C\sqrt{A}$ implies $A \leq B + C^2 + C\sqrt{B}$.)

4. Show that with probability $1 - \delta$, $\hat{h}$, the training error minimizer over $\mathcal{H}$, satisfies that
$$\mathrm{err}(\hat{h}, D) \leq \nu^\star + 6\sqrt{\frac{2 \ln \frac{2|\mathcal{H}|}{\delta}}{m} \nu^\star} + 8 \frac{\ln \frac{2|\mathcal{H}|}{\delta}}{m}.$$
(Hint: you may find the following elementary facts useful: for $A, B > 0$, $\sqrt{AB} \leq A + B$, $\sqrt{A + B} \leq \sqrt{A} + \sqrt{B}$. If you get other constants on the right hand side, no worries - you will still get full credit.)

5. Conclude that:
   (a) There exists a function m_A such that $m_A(\epsilon, \delta) = O(\frac{\ln |\mathcal{H}| + \ln \frac{1}{\delta}}{\epsilon^2})$, when $m \geq m_A(\epsilon, \delta)$, for all distributions D, $\mathrm{err}(\hat{h}, D) \leq \nu^\star + \epsilon$ with probability $1 - \delta$.
   (b) There exists a function m_R such that $m_R(\epsilon, \delta) = O(\frac{\ln |\mathcal{H}| + \ln \frac{1}{\delta}}{\epsilon})$, when $m \geq m_R(\epsilon, \delta)$, for all distributions D such that $\nu^\star = 0$, $\mathrm{err}(\hat{h}, D) \leq \epsilon$ with probability $1 - \delta$.