CSE512 Machine Learning

Chumki Acharya (112683478)

6 July, 2020

Homework-3

Question 1

Solution:

As per the given algorithm, we can see it maintains a set, V_t, of all the hypotheses which are consistent with $(x_1, y_1), ..., (x_{t-1}, y_{t-1})$. It then picks any hypothesis from V_t and predicts according to this hypothesis.

When it makes a prediction mistake, at least one hypothesis is removed from V_t.

Therefore, after making M mistakes we have

$$|V_t| \leq |\mathcal{H}| - M$$

Since V_t is always nonempty we have(by realizability assumption)

$$1 \leq |V_t| \leq |\mathcal{H}| - M$$

Rearranging, we obtain the following the mistake bound as

$$M(\mathcal{H}) \leq |\mathcal{H}| - 1$$

It is easy to construct a hypothesis class and a sequence of examples on which the algorithm will make $|\mathcal{H}| - 1$. mistakes.

Therefore, we can have a better algorithm in which we choose $h \in V_t$. We shall see that this algorithm is guaranteed to make exponentially fewer mistakes.

We simply note that whenever the algorithm errs we have

$$|V_{t+1}| \leq |V_t|/2$$

Therefore, if M is the total number of mistakes, we have

$$1 \leq |V_{T+1}| \leq |\mathcal{H}| 2^{-M}$$

Let $\mathcal{X} = \mathbb{R}^d$, and let $\mathcal{H} = \{h_1, \ldots, h_d\}$, where $h_j(\mathbf{x}) = \mathbb{1}_{(x_j=1]}$.

Let $\mathbf{x}_t = \mathbf{e}_t, y_t = \mathbb{1}_{[t=d]}, t = 1, \ldots, d$. The algorithm might predict $p_t = 1$ for every $t \in [d]$.

The number of mistakes done by the algorithm in this case is $d - 1 = |\mathcal{H}| - 1$.

Rearranging this we can conclude that it is not a strict inequality.

Question 2

Solution:

Let us use $G\left(C_1, \ldots, C_k\right)$ for the k -means objective, namely,

$$G\left(C_1, \ldots, C_k\right)=\min_{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k \in \mathbb{R}^n} \sum_{i=1}^k \sum_{\mathbf{x} \in C_i}\left\|\mathbf{x}-\boldsymbol{\mu}_i\right\|^2$$

Lets define $\boldsymbol{\mu}\left(C_i\right)=\frac{1}{\left|C_i\right|} \sum_{\mathbf{x} \in C_i} \mathbf{x}$ and note that $\boldsymbol{\mu}\left(C_i\right)=\operatorname{argmin}_{\boldsymbol{\mu} \in \mathbb{R}^n} \sum_{\mathbf{x} \in C_i}\|\mathbf{x}-\boldsymbol{\mu}\|^2$. Therefore, we can rewrite the k -means objective as

$$G\left(C_1, \ldots, C_k\right)=\sum_{i=1}^k \sum_{\mathbf{x} \in C_i}\left\|\mathbf{x}-\boldsymbol{\mu}\left(C_i\right)\right\|^2$$

Consider the update at iteration t of the k -means algorithm. Let $C_1^{(t-1)}, \ldots, C_k^{(t-1)}$ be the previous partition, let $\boldsymbol{\mu}_i^{(t-1)}=\boldsymbol{\mu}\left(C_i^{(t-1)}\right)$, and let $C_1^{(t)}, \ldots, C_k^{(t)}$ be the new partition assigned at iteration t. Using the definition of the objective as given in Equation (22.2) we clearly have that

$$G\left(C_1^{(t)}, \ldots, C_k^{(t)}\right) \leq \sum_{i=1}^k \sum_{\mathbf{x} \in C_i^{(t)}}\left\|\mathbf{x}-\boldsymbol{\mu}_i^{(t-1)}\right\|^2$$

In addition, the definition of the new partition $\left(C_1^{(t)}, \ldots, C_k^{(t)}\right)$ implies that it minimizes the expression $\sum_{i=1}^k \sum_{\mathbf{x} \in C_i}\left\|\mathbf{x}-\boldsymbol{\mu}_i^{(t-1)}\right\|^2$ over all possible partitions $\left(C_1, \ldots, C_k\right)$. Hence,

$$\sum_{i=1}^k \sum_{\mathbf{x} \in C_i^{(t)}}\left\|\mathbf{x}-\boldsymbol{\mu}_i^{(t-1)}\right\|^2 \leq \sum_{i=1}^k \sum_{\mathbf{x} \in C_i^{(t-1)}}\left\|\mathbf{x}-\boldsymbol{\mu}_i^{(t-1)}\right\|^2$$

Question 3

Solution:

Let us consider

$$\frac{\partial \ln p}{\partial \Sigma}=\frac{\partial}{\partial \Sigma} \sum_{n=1}^N \ln a_n=\sum_{n=1}^N \frac{1}{a_n} \frac{\partial a_n}{\partial \Sigma}.$$

We can define

$$a_n=\sum_{k=1}^K \pi_h \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma\right)$$

Again we can show

$$\frac{\partial \ln \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma\right)}{\partial \Sigma}=-\frac{1}{2} \Sigma^{-1}+\frac{1}{2} \Sigma^{-1} \mathbf{S}_{nh} \Sigma^{-1}$$

Where we have defined:

$$\mathbf{S}_{nk}=\left(\mathbf{x}_n-\boldsymbol{\mu}_k\right)\left(\mathbf{x}_n-\boldsymbol{\mu}_k\right)^T$$

Therefore, we can obtain:

$$
\begin{aligned}
\frac{\partial a_n}{\partial \Sigma} &= \frac{\partial}{\partial \Sigma}\left\{\sum_{k=1}^{K} \pi_k N\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}\right)\right\} \\
&= \sum_{k=1}^{K} \frac{\partial}{\partial \Sigma}\left\{\pi_k \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}\right)\right\} \\
&= \sum_{k=1}^{K} \pi_k \frac{\partial}{\partial \Sigma}\left\{\exp\left[\ln \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}\right)\right]\right\} \\
&= \sum_{k=1}^{K} \pi_k \cdot \exp\left[\ln \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}\right)\right] \cdot \frac{\partial}{\partial \Sigma}\left[\ln \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}\right)\right] \\
&= \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}\right) \cdot \left(-\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}_{nk}\boldsymbol{\Sigma}^{-1}\right)
\end{aligned}
$$

Substitute the equation above into we can obtain:

$$
\begin{aligned}
\frac{\partial \ln p}{\partial \Sigma} &= \sum_{n=1}^{N} \frac{1}{a_n}\frac{\partial a_n}{\partial \Sigma} \\
&= \sum_{n=1}^{N} \frac{\sum_{k=1}^{K} \pi_k \cdot N\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}\right) \cdot \left(-\frac{1}{2}\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}\mathbf{S}_{nk}\boldsymbol{\Sigma}^{-1}\right)}{\sum_{j=1}^{K} \pi_j \cdot N\left(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}\right)} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma\left(z_{nk}\right) \cdot \left(-\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}_{nk}\boldsymbol{\Sigma}^{-1}\right) \\
&= -\frac{1}{2}\left\{\sum_{n=1}^{N}\sum_{k=1}^{K} r\left(z_{nk}\right)\right\}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\boldsymbol{\Sigma}^{-1}\left\{\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma\left(z_{nk}\right)\mathbf{S}_{nk}\right\}\boldsymbol{\Sigma}^{-1}
\end{aligned}
$$

If we set the derivative equal to 0, we can obtain:

$$
\Sigma = \frac{\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma\left(z_{nk}\right)\mathbf{S}_{nk}}{\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma\left(z_{nk}\right)}
$$

Question 4

Solution:

We can show that $(\mathbb{E}[\hat{\sigma}])^2 = \frac{M-1}{M}\sigma^2$, thus we can conclude that σ is biased.

We can say that our proof holds for every random variable with finite variance.

Let $\mu = \mathbb{E}\left[x_1\right] = \ldots = \mathbb{E}\left[x_m\right]$ and let $\mu_2 = \mathbb{E}\left[x_1^2\right] = \ldots =$

$\mathbb{E}\left[x_m^2\right]$. Note that for $i \neq j, E\left[x_i x_j\right] = \mathbb{E}\left[x_i\right]\mathbb{E}\left[x_j\right] = \mu^2$

$$
\begin{aligned}
(\mathbb{E}[\hat{\sigma}])^2 &= \frac{1}{m}\sum_{i=1}^{m}\left(\mathbb{E}\left[x_i^2\right] - \frac{2}{m}\sum_{j=1}^{m}\mathbb{E}\left[x_i x_j\right] + \frac{1}{m^2}\sum_{j,k}\mathbb{E}\left[x_j x_k\right]\right) \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\mu_2 - \frac{2}{m}\left((m-1)\mu^2 + \mu_2\right) + \frac{1}{m^2}\left(m\mu_2 + m(m-1)\mu^2\right)\right) \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\frac{m-1}{m}\mu_2 - \frac{m-1}{m}\mu^2\right) \\
&= \frac{1}{m}\frac{m(m-1)}{m}\left(\mu_2 - \mu^2\right) \\
&= \frac{m-1}{m}\sigma^2
\end{aligned} \tag{1}
$$

Question 5

Solution:

Lets add one positive example and one negative example to the training sequence , denoted x_{m+1} and x_{m+2}, respectively.

Now we can see that the corresponding probabilities are θ and $1-\theta$. Hence, minimizing the RLM objective w.r.t. the original training sequence is equivalent to minimizing the ERM w.r.t. the extended training sequence. Therefore, the maximum likelihood estimator is given by

$$
\hat{\theta} = \frac{1}{m+2}\left(\sum_{i=1}^{m+2} x_i\right) = \frac{1}{m+2}\left(1 + \sum_{i=1}^{m} x_i\right) \tag{2}
$$

As per the hint given, we bound $\left|\hat{\theta} - \theta^\star\right|$ as -

$$
|\theta - \theta^\star| \leq |\hat{\theta} - \mathbb{E}[\hat{\theta}]| + \left|\mathbb{E}[\hat{\theta}] - \theta^\star\right|
$$

Further we bound each of the terms in the RHS of the last inequality For that we take,

$$
\mathbb{E}[\hat{\theta}] = \frac{1 + m\theta^\star}{m+2}
$$

Now, we have the following two inequalities.

$$
|\hat{\theta} - \mathbb{E}[\hat{\theta}]| = \frac{m}{m+2}\left|\frac{1}{m}\sum_{i=1}^{m} x_i - \theta^\star\right|
$$

$$
\left|\mathbb{E}[\hat{\theta}] - \theta^\star\right| = \frac{1 - 2\theta^\star}{m+2} \leq 1/(m+2)
$$

Applying Hoeffding's inequality, we obtain that for any $\epsilon > 0$,

$$
\mathbb{P}\left[|\theta - \theta^\star| \geq 1/(m+2) + \epsilon/2\right] \leq 2\exp\left(-m\epsilon^2/2\right)
$$

Thus, given a confidence parameter δ, the following bound holds with probability of at least

$$1-\delta' \, |\theta-\theta^{\star}| \leq O(\sqrt{\frac{\log(1/\delta)}{m}}) = \tilde{O}(1/\sqrt{m})$$

Question 6

Solution:

Let $\mathcal{X}=\mathbb{R}^d$ and we assume that each $\mathbf{x}$ is generated as follows. First, we choose a random number in $\{1,\ldots,k\}$. Let Y be a random variable corresponding to this choice, and denote $\mathcal{P}[Y=y]=c_y$. Second, we choose $\mathbf{x}$ on the basis of the value of Y according to a Gaussian distribution $\mathcal{P}[X=\mathbf{x} \mid Y=y]=\frac{1}{(2\pi)^{d/2}|\Sigma_y|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_y\right)^T \Sigma_y^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_y\right)\right)$

Let $Z_1,\ldots,Z_m$ be independent Bernoulli variables where for every $i, \mathbb{P}\left[Z_i=1\right]=p_i$ and $\mathbb{P}\left[Z_i=0\right]=1-p_i$. Let $p=\sum_{i=1}^m p_i$ and let $Z=\sum_{i=1}^m Z_i$. Using the monotonicity of the exponent function and Markov's inequality, we have that for every $t>0$

$$\begin{aligned}
\mathbb{P}[Z>(1+\delta)p]=\mathbb{P}\left[e^{tZ}>e^{t(1+\delta)p}\right] \leq \frac{\mathbb{E}\left[e^{tZ}\right]}{e^{(1+\delta)tp}} & \text{ by independence} \\
&=\prod_i \mathbb{E}\left[e^{tZ_i}\right] \\
&=\prod_i\left(p_i e^t+\left(1-p_i\right)e^0\right) \\
&=\prod_i\left(1+p_i\left(e^t-1\right)\right) \\
&\leq \prod_i e^{p_i\left(e^t-1\right)} \\
&=e^{\sum_i p_i\left(e^t-1\right)} \\
&=e^{\left(e^t-1\right)p}
\end{aligned}$$

Combining the above with the previous equation and choosing $t=\log(1+\delta)$ we obtain for every $i, \mathbb{P}\left[Z_i=1\right]=p_i$ and $\mathbb{P}\left[Z_i=0\right]=1-p_i$.

Let $p=\sum_{i=1}^m p_i$ and let $Z=\sum_{i=1}^m Z_i$ Then, for any $\delta>0$,

$$\mathbb{P}[Z>(1+\delta)p] \leq e^{-h(\delta)p}$$

using the inequality $h(a) \geq a^2/(2+2a/3)$

For the other direction, we apply similar calculations:

$$\mathbb{P}[Z<(1-\delta)p]=\mathbb{P}[-Z>-(1-\delta)p]=\mathbb{P}\left[e^{-tZ}>e^{-t(1-\delta)p}\right] \leq \frac{\mathbb{E}\left[e^{-tZ}\right]}{e^{-(1-\delta)tp}}, \log(Z)=(1+\delta)\log(1+\delta)-\delta$$

and,

$$\begin{aligned}
\mathbb{E}\left[e^{-tZ}\right] &= \mathbb{E}\left[e^{-t\sum_i Z_i}\right] = \mathbb{E}\left[\prod_i e^{-tZ_i}\right] \\
&= \prod_i \mathbb{E}\left[e^{-tz_i}\right] \\
&= \prod_i \left(1 + p_i\left(e^{-t} - 1\right)\right) \\
&\leq \prod_i e^{p_i\left(e^{-t}-1\right)} \\
&= e^{\left(e^{-t}-1\right)p}
\end{aligned}$$

Setting $t = -\log(1-\delta)$ yields

$$\mathbb{P}[Z < (1-\delta)p] \leq \frac{e^{-\delta p}}{e^{(1-\delta)\log(1-\delta)p}} = e^{-ph(-\delta)}$$

References

http://www.stat.cmu.edu/ larry/=sml2008/hw5_solution.pdf
http://www.stat.cmu.edu/Alon Gonen_Dana Rubinstein