# TTIC 31020: Introduction to Machine Learning
Problem Set 7

Hung Le Tran

20 Feb 2024

**Problem 7.1** (Problem 1)

**(a)**

$$y = \frac{3}{\sqrt{10}}x_{100} + \frac{1}{\sqrt{10}}x_1$$

where $x \sim N(0, \Sigma)$ with

$$\Sigma = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0.9 & \cdots & 0 \\ 0 & 0.9 & 1 & \cdots & 0 \\ & & \vdots & & \\ 0 & 0.9 & 0.9 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{100\times 100}$$

**1.** Optimal Feature Selection

Recall that

$$w_k = \arg\min_{\|w\|_0 \le k} L_S(w)$$

Then we'll have $I_1 = \{100\}$, $I_2 = \{1, 100\}$. There's no particular "order" as to which values are inserted into $I_k$, since for each $k$ we search from scratch. $w_2$ already achieves

$$L_S(w_2) = L_{\mathcal{D}}(w_2) = 0 \le 0.01$$

**2.** Greedy Feature Selection

$$I_{k+1} = \arg\min_I \min_{\text{supp}(w)\subseteq I} L_S(w) \text{ such that } I_k \subset I, |I| = k+1$$

Then $I_0 = \varnothing, I_1 = \{100\}, I_2 = \{100, 1\}$ with 100 getting added in the first iteration and 1 getting added in the second iteration. $w_2$ here also achieves 0 error.

**3.** $\ell_1$-norm relaxation.

Choose $B_1 = \frac{3}{\sqrt{10}}$ then $w_{B_1} = (0, 0, \ldots, 0, \frac{3}{\sqrt{10}})$ and $I_1 = \{100\}$.

Choose $B_2 = 1$ then $w_{B_2} = (\frac{1}{\sqrt{10}}, 0, \ldots, 0, \frac{3}{\sqrt{10}})$ and $I_2 = \{1, 100\}$. Fit $w_2$, then $w_2 = w_{B_2}$ and achieves 0 error.

**4.** Filter-by-correlation.

We have that

$$\mathbb{E}[y] = 0, Var(y) = \frac{9}{10}1 + \frac{1}{10}1 + 0 = 1$$

Therefore all the variables concerned $(x_1 \to x_{100}, y)$ have mean 0 and variance 1. We can calculate the correlation coefficients:

$$\rho_{100} = \frac{\text{Cov}(x_{100}, y)}{1} = \text{Cov}(x_{100}, \frac{3}{\sqrt{10}}x_{100} + \frac{1}{\sqrt{10}}x_1) = \frac{3}{\sqrt{10}}$$
$$\rho_1 = \frac{1}{\sqrt{10}}$$

then for $i$ with $2 \le i \le 99$, then

$$\begin{aligned}\rho_i &= \text{Cov}(x_i, \frac{3}{\sqrt{10}}x_{100} + \frac{1}{\sqrt{10}}x_1) \\ &= \frac{3}{\sqrt{10}}\text{Cov}(x_i, x_{100}) + \frac{1}{\sqrt{10}}\text{Cov}(x_i, x_1) \\ &= \frac{3}{\sqrt{10}}0.9 + 0 = \frac{2.7}{\sqrt{10}}\end{aligned}$$

Therefore $I_1 = \{100\}$, and $I_k$ for $k \in [2, 99]$ would include 100 and $(k-1)$ numbers in range $[2, 99]$ with arbitrary tie breaking. $I_{100} = [100]$ trivially.

But then for $I_k$, as $k$ increases, the inclusion of new features does not add any information to predict $y$. WLOG, we limit the prediction to

$$h_w(x) = w_{100}x_{100} + w_j x_j$$

for some $j \in [2, 99]$ for $k \le 99$. Then

$$\begin{aligned}L_S(w) &= \mathbb{E}[w_{100}x_{100} + w_j x_j - \frac{3}{\sqrt{10}}x_{100} - \frac{1}{\sqrt{10}}x_1] \\ &= \dots \\ &= \left(w_{100} - \frac{3}{\sqrt{10}}\right)^2 + w_j^2 + 0.9\left(w_{100} - \frac{3}{\sqrt{10}}\right)w_j + \frac{1}{10}\end{aligned}$$

hence the optimal weight would then be when

$$0 = \frac{\partial L}{\partial w_{100}} = \frac{\partial L}{\partial w_j}$$

which gives $w_{100} = \frac{3}{\sqrt{10}}, w_j = 0$, i.e., not using $x_j$ at all. This makes sense. Then the error would be

$$L_S(w) = 1/10$$

so we can't get any lower using $k \le 99$. Hence the smallest $k$ such that $L_{\mathcal{D}}(w_k) \le 0.01$ would be 100.

**5.** Conclusion: method 2 and 3 work as well as the optimal method.

**(b)**

$$x_1 = z_1, y = z_2, x_2 = z_1 + 0.0001z_2, x_i = z_i + 0.0001z_2$$

for $i \in \{3, 4, \dots, 100\}$, where $z \sim N(0, I)$.

**1.** Optimal Feature Selection.

$I_1$ is the singleton of any number in $[2, 100]$ with arbitrary tie breaking. WLOG, $I_1 = \{2\}$. Then $w = (w_2)$, and

$$\begin{aligned}L_S(w) &= \mathbb{E}[(w_2(z_1 + 0.0001z_2) - z_2)^2] \\ &= w_2^2 + (0.0001w_2 - 1)^2\end{aligned}$$

which has minimum of $\approx 0.9999$ at $w_2 = \frac{0.0002}{2+2\times10^{-8}} \approx 0.0001$. Our loss $0.9999 > 0.01$. Continue:

$I_2 = \{1, j\}$ for any $j \in [2, 100]$ with arbitrary tie breaking. WLOG, $j = 2$. Then $I_2 = \{1, 2\}$ with optimal weight $w = (-10^4, 10^4)$ achieving zero loss.

**2.** Greedy Feature Selection

Greedy selection selects $I_1 = \{j\}$ with arbitrary tie breaking for some $j \in [2, 100]$. This is because the best loss for $j \in [2, 100]$ would be $\approx 0.9999$, while if $I_1 = \{1\}$ then

$$\exp[w_1 x_1 - y] = \exp[w_1 z_1 - z_2] = w_1^2 + 1 \geq 1 > 0.9999$$

Then, $I_2 = \{j, 1\}$, with 1 added as the next feature, and the optimal weight is the aforementioned optimal weight. This weight achieves 0 loss.

**3.** $\ell_1$-norm relaxation.

Choose $B_1 = \frac{0.0002}{2+2\times 10^{-8}} \approx 0.0001$, then $w_{B_1}$ is the 1-sparse tuple containing $\frac{0.0002}{2+2\times 10^{-8}} \approx 0.0001$ at some index $j \in [2, 100]$. WLOG $j = 2$. Then $I_1 = \{2\}$. This weight, as aforementioned, achieves $\approx 0.9999$ loss.

Choose $B_2 = 2 \times 10^4$, then $w_{B_2}$ is 2-sparse with $10^{-4}$ at its first index and $10^4$ at some index $j \in [2, 100]$. WLOG $j = 2$, then $I_2 = \{1, 2\}$. This weight achieves 0 loss.

**4.** Filter-by-correlation.

We have trivially that $\rho_1 = 0$, while for $j \in [2, 100]$, say, $j = 2$, we have

$$\rho_2 = \frac{0.0001}{\sqrt{(1^2 + 0.0001^2)(1)}} \approx 9.9 \times 10^{-5}$$

Therefore, filter by correlation, for $k \in [99]$, would select $k$ numbers from $[2, 100]$ with arbitrary tie breaking, since $\rho_j \approx 9.9 \times 10^{-5} > 0 = \rho_1 \; \forall \; j \in [99]$.

However, $w_k$ would only be able to achieve $\approx 0.9999$ loss at best, therefore the smallest $k$ such that $L < 0.01$ would be 100, when $I_{100} = [100]$ trivially.

**5.** Conclusion: method 2 and 3 work as well as optimal feature selection.

## Problem 7.2

**(a)** WTS $L_{D^{(t+1)}}(h_t) = 0.5$.

We state the update rule for $D$:

$$D_i^{(t+1)} = \frac{D_i^{(t)} \exp(-\alpha_t y_i h_t(x_i))}{\sum_{j=1}^m D_j^{(t)} \exp(-\alpha_t y_j h_t(x_j))}$$

Let $E = \{i : h_t(x_i) \neq y_i\}$. Then for $i \in E$, we have

$$\begin{aligned} D_i^{(t)} \exp(-\alpha_t y_t h_t(x_i)) &= D_i^{(t)} \exp(\alpha_t) \\ &= D_i^{(t)} \left(\frac{1}{\varepsilon_t} - 1\right)^{1/2} \end{aligned}$$

and similarly if $i \notin E$ then

$$D_i^{(t)} \exp(-\alpha_t y_t h_t(x_i)) = D_i^{(t)} \left(\frac{1}{\varepsilon_t} - 1\right)^{-1/2}$$

It then follows that

$$\begin{aligned}
L_{D^{(t+1)}}(h_t) &= \sum D_i^{(t+1)} \mathbb{1}\{h_t(x_i) \neq y_i\} \\
&= \sum_{i \in E} D_i^{(t+1)} \\
&= \sum_{i \in E} \frac{D_i^{(t)} \left(\frac{1}{\varepsilon_t} - 1\right)^{1/2}}{\sum_{j \in E} D_j^{(t)} \exp(-\alpha_t y_i h_t(x_i))} \\
&= \frac{\sum_{j \in E} D_j^{(t)} \left(\frac{1}{\varepsilon_t} - 1\right)^{1/2}}{\sum_{j \in E} D_j^{(t)} \left(\frac{1}{\varepsilon_t} - 1\right)^{1/2} + \sum_{j \notin E} D_j^{(t)} \left(\frac{1}{\varepsilon_t} - 1\right)^{-1/2}} \\
&= \frac{\sum_{j \in E} D_j^{(t)} \left(\frac{1}{\varepsilon_t} - 1\right)}{\sum_{j \in E} D_j^{(t)} \left(\frac{1}{\varepsilon_t} - 1\right) + \sum_{j \notin E} D_j^{(t)}}
\end{aligned}$$

Recall that

$$\varepsilon_t = \sum_{j \in E} D_j^{(t)}$$

and

$$\sum_{j \notin E} D_j^{(t)} = 1 - \sum_{j \in E} D_j^{(t)}$$

so

$$\begin{aligned}
L_{D^{(t+1)}}(h_t) &= \frac{\sum_{j \in E} D_j^{(t)} (\sum_{j \in E} D_j^{(t)} - 1)}{\sum_{j \in E} D_j(t) (\sum_{j \in E} D_j^{(t)} - 1) - (1 - \sum_{j \in E} D_j^{(t)})(\sum_{j \in E} D_j^{(t)})} \\
&= \frac{a(a-1)}{a(a-1) - (1-a)a} = \frac{1}{2}
\end{aligned}$$

as required.

**(b)** We rewrite what we want to prove, using $T$ instead of $t$ to avoid confusion. WTS

$$\frac{\partial L_S^{\exp}(h_{w^{(T)}})}{\partial w[h]} \propto L_{D^{(T)}}^{01}(h) - \frac{1}{2}$$

We have that

$$h_{w^{(T)}}(x) = \sum_h w^{(T)}[h] \phi(x)[h] = \sum_h w^{(T)}[h] h(x)$$

which implies

$$\begin{aligned}
L_S^{\exp}(h_{w^{(T)}}) &= \frac{1}{m} \sum_{i=1}^m e^{-y_i h_{w^{(T)}}(x_i)} \\
&= \frac{1}{m} \sum_{i=1}^m e^{-y_i \sum_h w^{(T)}[h] h(x_i)} \\
\Rightarrow \frac{\partial L_S^{\exp}(h_{w^{(T)}})}{\partial w[h]} &= \frac{-1}{m} \sum_{i=1}^m \left[ y_i h(x_i) e^{-y_i \sum_h w^{(T)}[h] h(x_i)} \right] \\
&= \frac{-1}{m} \sum_{i=1}^m \left[ y_i h(x_i) e^{-y_i h_{w^{(T)}}(x_i)} \right] \qquad (1)
\end{aligned}$$

We now want to show that

$$\forall\, i \in [m], D_i^{(T)} = C_T \exp(-y_i h_{w^{(T)}}(x_i))$$

for some constant $C_T$ that only depends on $T$.

We will prove by induction on $T$.

For $T = 0$, for all $i \in [m]$, we have

$$D_i^{(0)} = \frac{1}{m}, \quad \exp(y_i h_{w^{(0)}(x_i)}) = e^0 = 1$$

so we have $C_0 = \frac{1}{m}$.

Suppose that the proposition holds for $T = T$, we now want to show that it is also true for $T = T + 1$. Indeed, for all $i \in [m]$, we have

$$\begin{aligned}
D_i^{(T+1)} &= \frac{D_i^{(T)} \exp(-\alpha_T y_i h_T(x_i))}{\sum_j D_j^{(T)} \exp(-\alpha_T y_j h_T(x_j))} \\
&= \frac{C_T}{C_T} \frac{\exp(-y_i h_{w^{(T)}}(x_i) - \alpha_T y_i h_T(x_i))}{\sum_j \exp(-y_j h_{w^{(T)}}(x_j) - \alpha_T y_j h_T(x_j))} \\
&= \frac{\exp(-y_i h_{w^{(T+1)}}(x_i))}{\sum_j \exp(-y_j h_{w^{(T+1)}}(x_j))} \\
&= C_{T+1} \exp(-y_i h_{w^{(T+1)}}(x_i))
\end{aligned}$$

where

$$C_{T+1} \coloneqq \frac{1}{\sum_j \exp(-y_j h_{w^{(T+1)}}(x_j))}$$

is only dependent on $T$.

By induction, we have that

$$D_i^{(T)} = C_T \exp(-y_i h_{w^{(T)}}(x_i))$$

holds for all $T$.

We now return to (1), recall that $E = \{i : h(x_i) \neq y_i\}$, and substitute $\exp(-y_i h_{w^{(T)}}(x_i)) = \frac{1}{C_T} D_i^{(T)}$, to have

$$\begin{aligned}
\frac{\partial L_S^{\exp}(h_{w^{(T)}})}{\partial w[h]} &= \frac{-1}{mC_T} \sum_{i=1}^{m} \left[ y_i h(x_i) D_i^{(T)} \right] \\
&= \frac{1}{mC_T} \left[ \sum_{i \in E} D_i^{(T)} - \sum_{i \notin E} D_i^{(T)} \right] \\
&= \frac{1}{C_T} \left[ L_{D^{(T)}}^{01}(h) - (1 - L_{D^{(T)}}^{01}(h)) \right] \\
&= \frac{2}{C_T} \left[ L_{D^{(T)}}^{01}(h) - \frac{1}{2} \right] \propto \left[ L_{D^{(T)}}^{01}(h) - \frac{1}{2} \right]
\end{aligned}$$

as required.