CMSC 25400 Homework 3

Ruolin Zheng

January 26, 2019

1 Problem 1

Let $\overrightarrow{\mathbf{y}} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$, $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$, $\mathbf{A}\boldsymbol{\theta} - \overrightarrow{\mathbf{y}} = \begin{bmatrix} h(\mathbf{x}_1) - y_1 \\ \vdots \\ h(\mathbf{x}_m) - y_m \end{bmatrix}$

$$
\begin{aligned}
J(\boldsymbol{\theta}) &= \frac{1}{2}\sum_{j=1}^{m}(h(\mathbf{x}_j) - y_j)^2 \\
&= \frac{1}{2}(\mathbf{A}\boldsymbol{\theta} - \overrightarrow{\mathbf{y}})^T(\mathbf{A}\boldsymbol{\theta} - \overrightarrow{\mathbf{y}}) \\
\nabla J(\boldsymbol{\theta}) &= \nabla(\frac{1}{2}(\mathbf{A}\boldsymbol{\theta} - \overrightarrow{\mathbf{y}})^T(\mathbf{A}\boldsymbol{\theta} - \overrightarrow{\mathbf{y}})) \\
&= \frac{1}{2}\nabla(\boldsymbol{\theta}^T\mathbf{A}^T\mathbf{A}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{A}^T\overrightarrow{\mathbf{y}} - \overrightarrow{\mathbf{y}}^T\mathbf{A}\boldsymbol{\theta} + \overrightarrow{\mathbf{y}}^T\overrightarrow{\mathbf{y}}) \\
&= \frac{1}{2}(\boldsymbol{\theta}^T(\mathbf{A}^T\mathbf{A} + (\mathbf{A}^T\mathbf{A})^T)) - \mathbf{A}^T\overrightarrow{\mathbf{y}} - (\overrightarrow{\mathbf{y}}^T\mathbf{A})^T) \\
&= \frac{1}{2}((\mathbf{A}^T\mathbf{A} + \mathbf{A}^T\mathbf{A})\boldsymbol{\theta} - \mathbf{A}^T\overrightarrow{\mathbf{y}} - \mathbf{A}^T\overrightarrow{\mathbf{y}}) \\
&= \frac{1}{2}(2\mathbf{A}^T\mathbf{A}\boldsymbol{\theta} - 2\mathbf{A}^T\overrightarrow{\mathbf{y}}) = 0 \\
\mathbf{A}^T\mathbf{A}\boldsymbol{\theta} &= \mathbf{A}^T\overrightarrow{\mathbf{y}} \\
\boldsymbol{\theta} &= (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\overrightarrow{\mathbf{y}}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{A}\boldsymbol{\theta} &= \begin{bmatrix} h(\mathbf{x}_1) \\ \vdots \\ h(\mathbf{x}_m) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n}\theta_i\phi_i(\mathbf{x}_1) \\ \vdots \\ \sum_{i=1}^{n}\theta_i\phi_i(\mathbf{x}_m) \end{bmatrix} = \begin{bmatrix} \theta_1\phi_1(\mathbf{x}_1) + \cdots + \theta_n\phi_n(\mathbf{x}_1) \\ \vdots \\ \theta_1\phi_1(\mathbf{x}_m) + \cdots + \theta_n\phi_n(\mathbf{x}_m) \end{bmatrix} \\
&= \begin{bmatrix} \phi_1(\mathbf{x}_1) & \dots & \phi_n(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_m) & \dots & \phi_n(\mathbf{x}_m) \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}
\end{aligned}
$$

Therefore, the design matrix $\mathbf{A}$ is as follows,

$$\mathbf{A} = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \dots & \phi_n(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_m) & \dots & \phi_n(\mathbf{x}_m) \end{bmatrix}$$

2 Problem 2

2.1 (a)

$$\begin{aligned}
l(\boldsymbol{\theta}) &= u \log\left(h(\mathbf{x})\right) + (1-u)\log\left(1-h(\mathbf{x})\right) \\
&= u \log\left(g(\boldsymbol{\theta}\cdot\mathbf{x})\right) + (1-u)\log\left(1-g(\boldsymbol{\theta}\cdot\mathbf{x})\right) \\
\nabla l(\boldsymbol{\theta}) &= u\frac{1}{g(\boldsymbol{\theta}\cdot\mathbf{x})}\nabla(g(\boldsymbol{\theta}\cdot\mathbf{x})) + (1-u)\frac{1}{1-g(\boldsymbol{\theta}\cdot\mathbf{x})}\nabla(-g(\boldsymbol{\theta}\cdot\mathbf{x})) \\
&= u\frac{1}{g(\boldsymbol{\theta}\cdot\mathbf{x})}g(\boldsymbol{\theta}\cdot\mathbf{x})(1-g(\boldsymbol{\theta}\cdot\mathbf{x}))\nabla(\boldsymbol{\theta}\cdot\mathbf{x}) \\
&\quad + (1-u)\frac{1}{1-g(\boldsymbol{\theta}\cdot\mathbf{x})}(-1)g(\boldsymbol{\theta}\cdot\mathbf{x})(1-g(\boldsymbol{\theta}\cdot\mathbf{x}))\nabla(\boldsymbol{\theta}\cdot\mathbf{x}) \\
&= u(1-g(\boldsymbol{\theta}\cdot\mathbf{x}))\mathbf{x} + (u-1)g(\boldsymbol{\theta}\cdot\mathbf{x})\mathbf{x} \\
&= u\mathbf{x} - u\cdot g(\boldsymbol{\theta}\cdot\mathbf{x}) + u\cdot g(\boldsymbol{\theta}\cdot\mathbf{x}) - g(\boldsymbol{\theta}\cdot\mathbf{x})\mathbf{x} \\
&= u\mathbf{x} - g(\boldsymbol{\theta}\cdot\mathbf{x})\mathbf{x} \\
&= (u - h(\mathbf{x}))\mathbf{x}
\end{aligned}$$

2.2 (b)

With $\nabla l(\boldsymbol{\theta}) = (u_i - h(\mathbf{x}_i))\mathbf{x}_i$, the SGD step on a single datapoint $(\mathbf{x}_i, u_i)$ is as follows,

$$\begin{aligned}
\boldsymbol{\theta} &\leftarrow \boldsymbol{\theta} + \alpha[\nabla l(\boldsymbol{\theta})] \\
&= \boldsymbol{\theta} + \alpha[(u_i - h(\mathbf{x}_i))\mathbf{x}_i] \\
&= \boldsymbol{\theta} - \alpha[(h(\mathbf{x}_i) - u_i)\mathbf{x}_i]
\end{aligned}$$

3 Problem 3

Proof. For a given data set, let $U_A = \{u_1, \dots, u_k\}$ denote the sequence of updates A makes to its hypothesis, $|U_A| \leq M$. Suppose there is a conservative algorithm A' which runs on the same data set. For each update u_i made by A, A' will make the same update to its hypothesis if it makes a mistake, and otherwise will skip that update. This implies that some $u_i \in U_A$ may not be in $U_{A'}$.
Therefore, $U_{A'} \subseteq U_A \implies |U_{A'}| \leq |U_A| \leq M$. This proves the claim that there is a conservative algorithm A' for $\mathcal{C}$ which also has a finite mistake bound M. □

4 Problem 4

4.1 (a)

Proof. In the $k = 2$ case,

$$\mathbf{v}_{y_t} \cdot \mathbf{x}_t - \mathbf{v}_{\overline{y}} \cdot \mathbf{x}_t \geq 2\delta \qquad \overline{y} \in \{1, 2\} \setminus \{y_t\}$$

The definition of the margin we saw in class is as follows: assume the data is separable with margin δ, let unit vector $\mathbf{v}$ satisfy $|\mathbf{v} \cdot \mathbf{x}_t| \geq \delta$ for all $\mathbf{x}_t$. Here, $\mathbf{v}$ is the normal vector of the separating hyperplane. Any $\mathbf{x}_t$ is predicted to be of class 1 if it points in the same direction as $\mathbf{v}$ and class 0 if it points in the opposite direction.
In the definition above where we have unit vectors $\mathbf{v}_1, \mathbf{v}_2$, let $\mathbf{v}_1 = \mathbf{v}$ and $\mathbf{v}_2 = -\mathbf{v}$. Any $\mathbf{x}_t$ is predicted to be of class 1 if it points in the same direction as $\mathbf{v}_1$ and class 2 if it points in the same direction as $\mathbf{v}_2$, opposite to $\mathbf{v}_1$.
This gives us,

$$\begin{aligned}
\mathbf{v}_1 \cdot \mathbf{x}_t = \mathbf{v} \cdot \mathbf{x}_t &\geq \delta \\
-\mathbf{v}_2 \cdot \mathbf{x}_t = \mathbf{v} \cdot \mathbf{x}_t &\geq \delta \\
\mathbf{v}_1 \cdot \mathbf{x}_t - \mathbf{v}_2 \cdot \mathbf{x}_t = \mathbf{v} \cdot \mathbf{x}_t + \mathbf{v} \cdot \mathbf{x}_t &\geq \delta + \delta \\
2\mathbf{v} \cdot \mathbf{x}_t &\geq 2\delta \\
\mathbf{v} \cdot \mathbf{x}_t &\geq \delta
\end{aligned}$$

Therefore, when $k = 2$, $\mathbf{v}_{y_t} \cdot \mathbf{x}_t - \mathbf{v}_{\overline{y}} \cdot \mathbf{x}_t \geq 2\delta, \overline{y} \in \{1, 2\} \setminus \{y_t\}$ is equivalent to $|\mathbf{v} \cdot \mathbf{x}_t| \geq \delta$ which we saw in class. □

4.2 (b)

Let $a = \mathbf{w}_1 \cdot \mathbf{v}_1 + \mathbf{w}_2 \cdot \mathbf{v}_2 + \mathbf{w}_3 \cdot \mathbf{v}_3, b = ||\mathbf{w}_1||^2 + ||\mathbf{w}_2||^2 + ||\mathbf{w}_3||^2$.
Claim 1. After M mistakes, a increases by at least δM.

Proof. Initially, $\mathbf{w}_1 \cdot \mathbf{v}_1 = \mathbf{w}_2 \cdot \mathbf{v}_2 = \mathbf{w}_3 \cdot \mathbf{v}_3 = 0$.
WLOG, consider the case of mistake in which we wrongly predicted a datapoint of class 1 to be in class 2. That is, $y_t = 1$ but $\hat{y}_t = 2$.
The update rules are as follows,

$$\begin{aligned}
\mathbf{w}_1 &\leftarrow \mathbf{w}_1 + \mathbf{x}/2 \\
\mathbf{w}_2 &\leftarrow \mathbf{w}_2 - \mathbf{x}/2 \\
\mathbf{w}_1 \cdot \mathbf{v}_1 &\leftarrow (\mathbf{w}_1 + \mathbf{x}/2) \cdot \mathbf{v}_1 = \mathbf{w}_1 \cdot \mathbf{v}_1 + \mathbf{v}_1 \cdot \mathbf{x}/2 \\
\mathbf{w}_2 \cdot \mathbf{v}_2 &\leftarrow (\mathbf{w}_2 - \mathbf{x}/2) \cdot \mathbf{v}_2 = \mathbf{w}_2 \cdot \mathbf{v}_2 - \mathbf{v}_2 \cdot \mathbf{x}/2
\end{aligned}$$

Given that $\mathbf{v}_{y_t} \cdot \mathbf{x} - \mathbf{v}_{\overline{y}} \cdot \mathbf{x} \geq 2\delta \quad \overline{y} \in \{1, 2, 3\} \setminus \{y_t\}$,

$$\begin{aligned}
\mathbf{v}_1 \cdot \mathbf{x} - \mathbf{v}_2 \cdot \mathbf{x} &\geq 2\delta \\
\mathbf{v}_1 \cdot \mathbf{x}/2 - \mathbf{v}_2 \cdot \mathbf{x}/2 &\geq \delta
\end{aligned}$$

Therefore, the following holds for any $(\mathbf{x}, y_t)$, where $y_t, \hat{y}_t \in \{1, 2, 3\}$ and $\hat{y}_t \neq y_t$.

$$\begin{aligned} a_{new} = \mathbf{w}_1 \cdot \mathbf{x} + \mathbf{w}_2 \cdot \mathbf{x} + \mathbf{w}_2 \cdot \mathbf{x} + \mathbf{w}_3 \cdot \mathbf{x} &\leftarrow \mathbf{w}_1 \cdot \mathbf{x} + \mathbf{w}_2 \cdot \mathbf{x} + \mathbf{w}_2 \cdot \mathbf{x} + \mathbf{w}_3 \cdot \mathbf{x} \\ &\quad + \mathbf{v}_1 \cdot \mathbf{x}/2 - \mathbf{v}_2 \cdot \mathbf{x}/2 \\ &\geq a + \delta \end{aligned}$$

Every time the algorithm makes a mistake, a increases by at least δ. Therefore, after M mistakes, a increases by at least δM. □

Claim 2. After M mistakes, b increases by at most $\frac{1}{2}M$, recall that $b = ||\mathbf{w}_1||^2 + ||\mathbf{w}_2||^2 + ||\mathbf{w}_3||^2$.

Proof. Similar as above, WOLG, consider the case of mistake in which $y_t = 1$ but $\hat{y}_t = 2$. The updates are as follows,

$$\begin{aligned} ||\mathbf{w}_1||^2 &\leftarrow (\mathbf{w}_1 + \mathbf{x}/2) \cdot (\mathbf{w}_1 + \mathbf{x}/2) = ||\mathbf{w}_1||^2 + 2(\mathbf{w}_1 \cdot \mathbf{x}/2) + ||\mathbf{x}/2||^2 \\ &= ||\mathbf{w}_1||^2 + 2(\mathbf{w}_1 \cdot \mathbf{x}/2) + \frac{1}{2} \\ ||\mathbf{w}_2||^2 &\leftarrow (\mathbf{w}_2 - \mathbf{x}/2) \cdot (\mathbf{w}_2 - \mathbf{x}/2) = ||\mathbf{w}_2||^2 - 2(\mathbf{w}_2 \cdot \mathbf{x}/2) + ||\mathbf{x}/2||^2 \\ &= ||\mathbf{w}_2||^2 - 2(\mathbf{w}_2 \cdot \mathbf{x}/2) + \frac{1}{2} \end{aligned}$$

Because we predicted the datapoint to be in class 2, $\hat{y} = 2$, according to our prediction rule $\hat{y} = \text{argmax}_{i \in \{1,\ldots,k\}}(\mathbf{w}_i \cdot x)$, we have,

$$\begin{aligned} \mathbf{w}_2 \cdot x &\geq \mathbf{w}_1 \cdot x \\ \mathbf{w}_1 \cdot x - \mathbf{w}_2 \cdot x &\leq 0 \end{aligned}$$

We also assume $||\mathbf{x}|| = 1$ and thus $||\mathbf{x}/2|| = \frac{1}{2}$. Therefore, the following holds for any $(\mathbf{x}, y_t)$, where $y_t, \hat{y}_t \in \{1, 2, 3\}$ and $\hat{y}_t \neq y_t$.

$$\begin{aligned} b_{new} = ||\mathbf{w}_1||^2 + ||\mathbf{w}_2||^2 + ||\mathbf{w}_3||^2 &\leftarrow ||\mathbf{w}_1||^2 + ||\mathbf{w}_2||^2 + ||\mathbf{w}_3||^2 \\ &\quad + 2(\mathbf{w}_1 \cdot \mathbf{x}/2 - \mathbf{w}_2 \cdot \mathbf{x}/2) + 2||\mathbf{x}/2||^2 \\ &\leq b + 0 + 2 \cdot \frac{1}{4} = b + \frac{1}{2} \end{aligned}$$

Every time the algorithm makes a mistake, b increases by at most $\frac{1}{2}$. Therefore, after M mistakes, b increases by at most $\frac{1}{2}M$. □

We now have $a = \mathbf{w}_1 \cdot \mathbf{v}_1 + \mathbf{w}_2 \cdot \mathbf{v}_2 + \mathbf{w}_3 \cdot \mathbf{v}_3 \geq \delta M, b = ||\mathbf{w}_1||^2 + ||\mathbf{w}_2||^2 + ||\mathbf{w}_3||^2 \leq \frac{1}{2}M$. From here we derive an inequality that involves a and b.

$$\begin{aligned} a &= \mathbf{w}_1 \cdot \mathbf{v}_1 + \mathbf{w}_2 \cdot \mathbf{v}_2 + \mathbf{w}_3 \cdot \mathbf{v}_3 \\ &\leq ||\mathbf{w}_1|| + ||\mathbf{w}_2|| + ||\mathbf{w}_3|| \\ &= \sqrt{(||\mathbf{w}_1|| + ||\mathbf{w}_2|| + ||\mathbf{w}_3||)^2} \\ &= \sqrt{||\mathbf{w}_1||^2 + ||\mathbf{w}_2||^2 + ||\mathbf{w}_3||^2 + 2||\mathbf{w}_1||||\mathbf{w}_2|| + 2||\mathbf{w}_2||||\mathbf{w}_3|| + 2||\mathbf{w}_1||||\mathbf{w}_3||} \end{aligned}$$

It follows from $(x-y)^2 = x^2 + y^2 - 2xy \geq 0, x^2 + y^2 \geq 2xy$ that $2||\mathbf{w}_i||||\mathbf{w}_j|| \leq ||\mathbf{w}_i||^2$ for any i, j.

$$\begin{aligned} a &\leq \sqrt{||\mathbf{w}_1||^2 + ||\mathbf{w}_2||^2 + ||\mathbf{w}_3||^2 + 2||\mathbf{w}_1||^2 + 2||\mathbf{w}_2||^2 + 2||\mathbf{w}_3||^2} \\ &= \sqrt{3(||\mathbf{w}_1||^2 + ||\mathbf{w}_2||^2 + ||\mathbf{w}_3||^2)} \\ &= \sqrt{3b} \end{aligned}$$

Recall $a \geq \delta M, b \leq \frac{1}{2}M$. Therefore,

$$\begin{aligned} \delta M \leq a \leq \sqrt{3b} &\leq \sqrt{3 \cdot \frac{1}{2}M} \\ \delta^2 M^2 &\leq 3 \cdot \frac{1}{2}M \\ M &\leq \frac{3}{2} \cdot \frac{1}{\delta^2} \end{aligned}$$

The mistake bound M is thus $\frac{3}{2} \cdot \frac{1}{\delta^2}$.

5 Problem 5

Comment. With a 10-fold cross validation, the sequence of error estimates produced by $M = 1, \ldots, 10$ is $[0.106, 0.075, 0.0705, 0.065, 0.074, 0.0735, 0.0625, 0.066, 0.0645, 0.0595]$. The mistakes frequency reaches the smallest value 0.0595 at $M = 10$, so we feed the data set to the Perceptron Algorithm 10 times. The plots of mistakes when the data set is fed the first time as well as over all 10 runs are shown below.
It is especially evident in Figure 2 that the graph has a concave shape. As expected, as the number of examples seen increases, the cumulative number of mistakes grows at a decreasing rate.

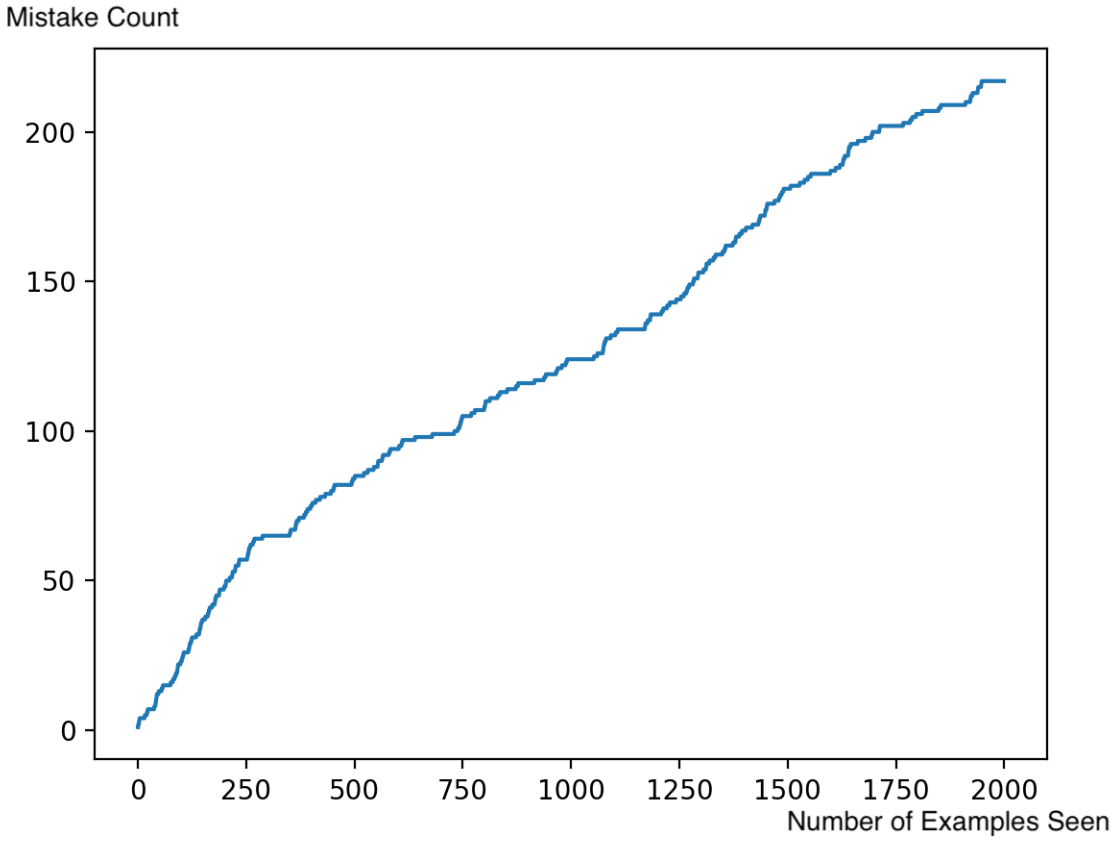

Figure 1: Number of Mistakes When the Data Set is Fed the First Time

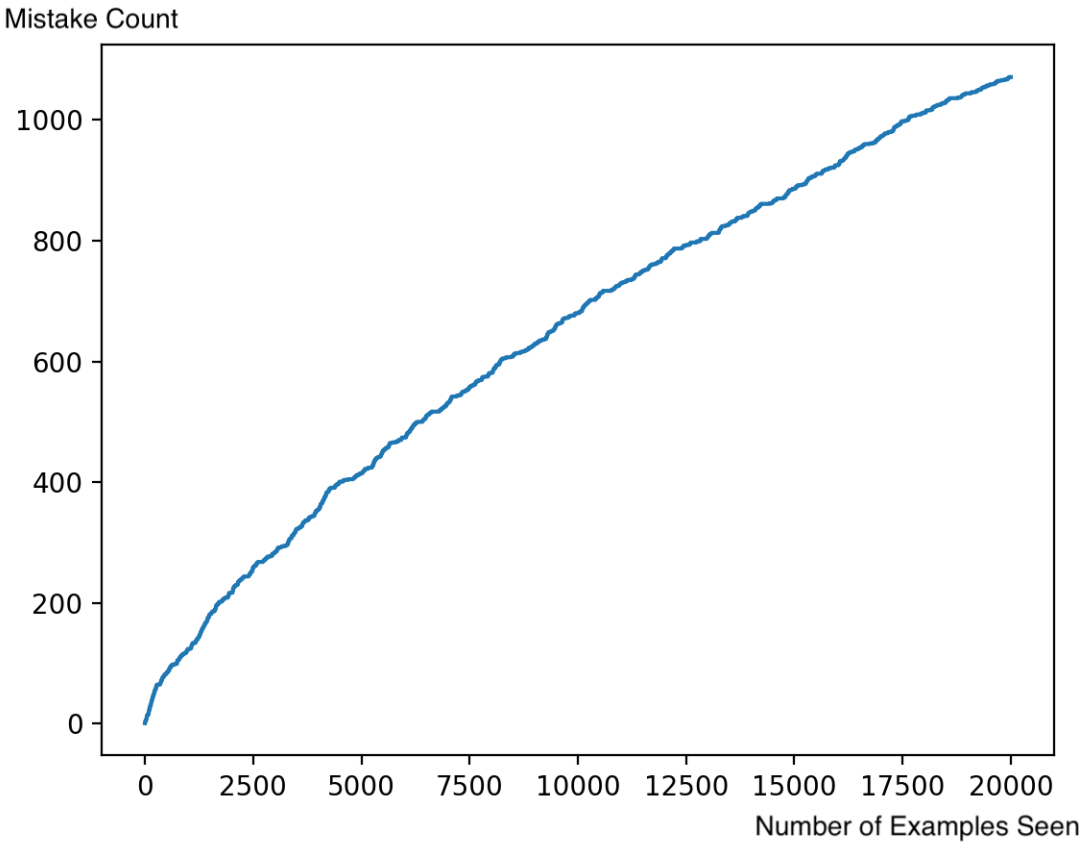

Figure 2: Number of Mistakes When the Data Set is Fed 10 Times

Comment (Continued): Below is a plot of 20 randomly-chosen test datapoints and the corresponding predictions made by the algorithm. It appears from the plot that the result is quite good. Predictions for the entire *test35.digits* set can be found in the file *test35.predictions*.

Figure 3: 20 Randomly-Chosen Test datapoints and Corresponding Predictions