Introduction to Machine Learning
Fall 2019
University of Science and Technology of China

Lecturer: Jie Wang | Homework 3
Posted: Oct. 21, 2019 | Due: Oct. 30, 2019
Name: San Zhang | ID: PBXXXXXXXX

Notice, to get the full credits, please present your solutions step by step.

Exercise 1: Logistic Regression 40pts

Given the training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$. Let

$$\begin{aligned}\mathcal{I}^+ =& \{i : i \in [n], y_i = 1\},\\ \mathcal{I}^- =& \{i : i \in [n], y_i = 0\},\end{aligned}$$

where $[n] = \{1, 2, \ldots, n\}$. We assume that $\mathcal{I}^+$ and $\mathcal{I}^-$ are not empty.

Then, we can formulate the logistic regression as:

$$\min_{\mathbf{w}} \; L(\mathbf{w}) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log(\frac{\exp(\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle)}{1+\exp(\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle)}) + (1-y_i)\log(\frac{1}{1+\exp(\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle)})\right), \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^{d+1}$ is the model parameter to be estimated and $\overline{\mathbf{x}}_i^\top = (1, \mathbf{x}_i^\top)$.

1. Find the gradient and the Hessian of $L(\mathbf{w})$.

2. Suppose that $\overline{\mathbf{X}} = (\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, \ldots, \overline{\mathbf{x}}_n)^\top \in \mathbb{R}^{n\times(d+1)}$ and $\mathbf{rank}(\overline{\mathbf{X}}) = d + 1$. Show that $L(\mathbf{w})$ is strictly convex, i.e., for all $\mathbf{w}_1 \neq \mathbf{w}_2$,

$$L(t\mathbf{w}_1 + (1-t)\mathbf{w}_2) < tL(\mathbf{w}_1) + (1-t)L(\mathbf{w}_2), \forall\, t \in (0, 1).$$

3. Suppose that the training data is linearly separable, that is, there exists $\hat{\mathbf{w}} \in \mathbb{R}^{d+1}$ such that

$$\begin{aligned}&\langle \hat{\mathbf{w}}, \overline{\mathbf{x}}_i\rangle > 0, \forall\, i \in \mathcal{I}^+,\\ &\langle \hat{\mathbf{w}}, \overline{\mathbf{x}}_i\rangle < 0, \forall\, i \in \mathcal{I}^-.\end{aligned}$$

Show that problem (1) has no solution.

4. **(Bonus** 20pts) Suppose that the training data is NOT linearly separable. Show that problem (1) always admits a solution. Moreover, show that the solution is unique.

Solution: 1. The gradient of $L(\mathbf{w})$ is

$$\begin{aligned}\nabla L(\mathbf{w}) &= -\frac{1}{n}\sum_{i=1}^{n}(y_i \cdot \frac{1}{1+\exp(\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle)} \cdot \overline{\mathbf{x}_i} - (1-y_i)\cdot\frac{\exp(\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle)}{1+\exp(\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle)} \cdot \overline{\mathbf{x}_i})\\ &= \frac{1}{n}\sum_{i=1}^{n}(\frac{1}{1+\exp(-\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle)} - y_i)\overline{\mathbf{x}}_i.\end{aligned}$$

The Hessian of $L(\mathbf{w})$ is

$$\begin{aligned}\nabla^2 L(\mathbf{w}) &= \frac{1}{n}\sum_{i=1}^{n}(y_i \cdot \frac{\exp(\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle)}{(1+\exp(\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle))^2} \cdot \overline{\mathbf{x}_i} \cdot \overline{\mathbf{x}_i}^\top + (1-y_i) \cdot \frac{\exp(\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle)}{(1+\exp(\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle))^2} \cdot \overline{\mathbf{x}_i} \cdot \overline{\mathbf{x}_i}^\top \\ &= \frac{1}{n}\sum_{i=1}^{n}(\frac{\exp(\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle)}{(1+\exp(\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle))^2} \cdot \overline{\mathbf{x}_i} \cdot \overline{\mathbf{x}_i}^\top).\end{aligned}$$

2. Consider the functions $f_+(x) = \log(1+\exp(-x))$ and $f_-(x) = \log(1+\exp(x))$. Note that both $f_+(x)$ and $f_-(x)$ are strictly convex since

$$f_+''(x) = \frac{\exp(-x)}{(1+\exp(-x))^2} > 0 \text{ and } f_-''(x) = \frac{\exp(x)}{(1+\exp(x))^2} > 0.$$

For all $\mathbf{w}_1 \neq \mathbf{w}_2$, there exists $j \in [n]$ such that $\langle \mathbf{w}_1, \overline{\mathbf{x}}_j\rangle \neq \langle \mathbf{w}_2, \overline{\mathbf{x}}_j\rangle$ since $\overline{\mathbf{X}}^\top(\mathbf{w}_1 - \mathbf{w}_2) \neq \mathbf{0}$

Let $\mathbf{w}_3 = t\mathbf{w}_1 + (1-t)\mathbf{w}_2$ with $t \in (0,1)$. Then,

$$\begin{aligned}L(\mathbf{w}_3) &= \frac{1}{n}\sum_{i=1}^{n} y_i f_+(\langle \mathbf{w}_3, \overline{\mathbf{x}}_i\rangle) + (1-y_i) f_-(\langle \mathbf{w}_3, \overline{\mathbf{x}}_i\rangle) \\ &= \frac{1}{n}\,(y_j f_+(\langle \mathbf{w}_3, \overline{\mathbf{x}}_j\rangle) + (1-y_j) f_-(\langle \mathbf{w}_3, \overline{\mathbf{x}}_j\rangle) \\ &\quad + \sum_{i\neq j} y_i f_+(\langle \mathbf{w}_3, \overline{\mathbf{x}}_i\rangle) + (1-y_i) f_-(\langle \mathbf{w}_3, \overline{\mathbf{x}}_i\rangle).\end{aligned}$$

Since f^+, f^- are strongly convex, we have

$$\begin{aligned}&f_+(\langle \mathbf{w}_3, \overline{\mathbf{x}}_i\rangle) \leq t f_+(\langle \mathbf{w}_1, \overline{\mathbf{x}}_i\rangle) + (1-t) f_+(\langle \mathbf{w}_2, \overline{\mathbf{x}}_i\rangle) \text{ if } i \neq j, \\ &f_-(\langle \mathbf{w}_3, \overline{\mathbf{x}}_i\rangle) \leq t f_-(\langle \mathbf{w}_1, \overline{\mathbf{x}}_i\rangle) + (1-t) f_-(\langle \mathbf{w}_2, \overline{\mathbf{x}}_i\rangle) \text{ if } i \neq j, \\ &f_+(\langle \mathbf{w}_3, \overline{\mathbf{x}}_i\rangle) < t f_+(\langle \mathbf{w}_1, \overline{\mathbf{x}}_i\rangle) + (1-t) f_+(\langle \mathbf{w}_2, \overline{\mathbf{x}}_i\rangle) \text{ if } i = j, \quad (\langle \mathbf{w}_1, \overline{\mathbf{x}}_j\rangle \neq \langle \mathbf{w}_2, \overline{\mathbf{x}}_j\rangle) \\ &f_-(\langle \mathbf{w}_3, \overline{\mathbf{x}}_i\rangle) < t f_-(\langle \mathbf{w}_1, \overline{\mathbf{x}}_i\rangle) + (1-t) f_-(\langle \mathbf{w}_2, \overline{\mathbf{x}}_i\rangle) \text{ if } i = j. \quad (\langle \mathbf{w}_1, \overline{\mathbf{x}}_j\rangle \neq \langle \mathbf{w}_2, \overline{\mathbf{x}}_j\rangle)\end{aligned}$$

Therefore,

$$L(\mathbf{w}_3) < tL(\mathbf{w}_1) + (1-t)L(\mathbf{w}_2).$$

3. Since the training data is linearly separable, there exists $\hat{\mathbf{w}} \in \mathbb{R}^{d+1}$ such that

$$\begin{aligned}\langle \hat{\mathbf{w}}, \overline{\mathbf{x}}_i\rangle > 0, \forall\, i \in \mathcal{I}^+, \\ \langle \hat{\mathbf{w}}, \overline{\mathbf{x}}_i\rangle < 0, \forall\, i \in \mathcal{I}^-.\end{aligned}$$

Consider $\lambda\hat{\mathbf{w}}$ with $\lambda > 0$. Then,

$$\begin{aligned}\lim_{\lambda\to+\infty} \langle \lambda\hat{\mathbf{w}}, \overline{\mathbf{x}}_i\rangle = +\infty, \forall\, i \in \mathcal{I}^+, \\ \lim_{\lambda\to+\infty} \langle \lambda\hat{\mathbf{w}}, \overline{\mathbf{x}}_i\rangle = -\infty, \forall\, i \in \mathcal{I}^-.\end{aligned}$$

$$
\begin{aligned}
\lim_{\lambda\to+\infty} L(\lambda\hat{\mathbf{w}}) &= \lim_{\lambda\to+\infty} \frac{1}{n}\left(\sum_{i\in\mathcal{L}^+} f_+(\lambda\langle\hat{\mathbf{w}},\overline{\mathbf{x}}_i\rangle) + \sum_{i\in\mathcal{L}^-} f_-(\lambda\langle\hat{\mathbf{w}},\overline{\mathbf{x}}_i\rangle)\right)\\
&= 0\\
\Rightarrow \inf_{\mathbf{w}} L(\mathbf{w}) &\le 0.
\end{aligned}
$$

Let $\mathbf{w}^*$ be a solution to (1). Thus

$$
\begin{aligned}
L(\mathbf{w}^*) &\le 0\\
\Rightarrow \frac{\exp(\langle\mathbf{w}^*,\overline{\mathbf{x}}_i\rangle)}{\exp(\langle\mathbf{w}^*,\overline{\mathbf{x}}_i\rangle)+1} &= 0,\ \forall i = 1,\ldots,n.
\end{aligned}
\tag{2}
$$

Therefore, the equation (2) has no solution and hence the problem (1) has no solution.

4. The training data is NOT linearly separable, that is,

$$
\forall\, \mathbf{w}\in\mathbb{R}^{d+1}, \exists\, i\in[n] \quad \text{s.t.} \quad \begin{cases} \langle\mathbf{w},\overline{\mathbf{x}}_i\rangle < 0 & \text{if } y_i = 1,\\ \langle\mathbf{w},\overline{\mathbf{x}}_i\rangle > 0 & \text{if } y_i = 0. \end{cases} \tag{3}
$$

Note that the inverse proposition of (3) does not imply the training data is linearly separable.

Suppose $\mathbf{rank}(\overline{\mathbf{X}}) = r \le d+1$. Let $[\,\cdot\,]_{1:r}$ denote the first r entries of a vector.

WLOG, we assume that $\overline{\mathbf{X}}_r = ([\overline{\mathbf{x}}_1]_{1:r} \quad \ldots \quad [\overline{\mathbf{x}}_r]_{1:r})^\top$ and $\mathbf{rank}(\overline{\mathbf{X}}_r) = r$. Thus $\overline{\mathbf{X}} = \overline{\mathbf{X}}_r(\mathbf{I}_r \quad \mathbf{M})$, where $\mathbf{M}\in\mathbf{R}^{r\times(d+1-r)}$. Let $\mathbf{u} = (\mathbf{I}_r \quad \mathbf{M})\mathbf{w}\in\mathbb{R}^r$. Define

$$
L^*(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^n y_i f_+(\langle\mathbf{u},[\overline{\mathbf{x}}_i]_{1:r}\rangle) + (1-y_i) f_-(\langle\mathbf{u},[\overline{\mathbf{x}}_i]_{1:r}\rangle).
$$

Note that $L^*((\mathbf{I}_r \quad \mathbf{M})\mathbf{w}) = L(\mathbf{w})$ and $\{([\mathbf{x}_i]_{1:r}, y_i)\}_{i=1}^n$ is not linearly separable.

Now we show that $L^*(\mathbf{u})$ always admits a solution by following steps.

(a) Firstly, we can define a continuous function $g_{\max}(\mathbf{u})$ that satisfies $g_{\max}(\mathbf{u}) < L^*(\mathbf{u})$ for all $\mathbf{u}\in\mathbb{R}^r$.

Define

$$
g_i(\mathbf{u}) = \frac{1}{n}\left(y_i f_+(\langle\mathbf{u},[\overline{\mathbf{x}}_i]_{1:r}\rangle) + (1-y_i) f_-(\langle\mathbf{u},[\overline{\mathbf{x}}_i]_{1:r}\rangle)\right),
$$

and

$$
g_{\max}(\mathbf{u}) = \max_{i\in[n]} g_i(\mathbf{u}).
$$

Note that $g_{\max}(\mathbf{u})$ is continuous, and $g_{\max}(\mathbf{u}) < L^*(\mathbf{u})$ for all $\mathbf{u}$.

(b) For any $u \in \mathbb{R}^r$, we can find some $i_{\mathbf{u}} \in [n]$ such that $g_{\max}(\mathbf{u}) = g_{i_{\mathbf{u}}}(\mathbf{u})$.
Since $\{([\mathbf{x}_i]_{1:r}, y_i)\}_{i=1}^n$ is not linearly separable for any $u \in \mathbb{R}^r$, $i_{\mathbf{u}}$ must satisfies

$$\langle \mathbf{u}, [\overline{\mathbf{x}}_{i_{\mathbf{u}}}]_{1:r} \rangle < 0, \text{ if } i_{\mathbf{u}} \in \mathcal{L}^+;$$
$$\langle \mathbf{u}, [\overline{\mathbf{x}}_{i_{\mathbf{u}}}]_{1:r} \rangle > 0, \text{ if } i_{\mathbf{u}} \in \mathcal{L}^-.$$

Further, if $g_{\max}(\mathbf{u}) = g_{i_{\mathbf{u}}}(\mathbf{u})$ and $\alpha > 0$, then $g_{\max}(\alpha\mathbf{u}) = g_{i_{\alpha\mathbf{u}}}(\alpha\mathbf{u}) = g_{i_{\mathbf{u}}}(\alpha\mathbf{u})$ for all $\alpha > 0$, that is because

$$\begin{aligned} g_{i_{\mathbf{u}}}(\alpha\mathbf{u}) &= \frac{1}{n}\left(y_{i_{\mathbf{u}}} f_+(\langle \alpha\mathbf{u}, [\overline{\mathbf{x}}_{i_{\mathbf{u}}}]_{1:r} \rangle) + (1 - y_{i_{\mathbf{u}}}) f_-(\langle \alpha\mathbf{u}, [\overline{\mathbf{x}}_{i_{\mathbf{u}}}]_{1:r} \rangle)\right) \\ &\geq \frac{1}{n}\left(y_j f_+(\langle \alpha\mathbf{u}, [\overline{\mathbf{x}}_j]_{1:r} \rangle) + (1 - y_j) f_-(\langle \alpha\mathbf{u}, [\overline{\mathbf{x}}_j]_{1:r} \rangle)\right) \\ &= g_j(\alpha\mathbf{u}), \forall\, j \in [n]\,. \end{aligned}$$

As $g_{i_{\mathbf{u}}}(\alpha\mathbf{u})$ is an increasing function with respect to α ($\alpha > 0$), we have $g_{\max}(\alpha\mathbf{u})$ is also an increasing function with respect to α. Moreover,

$$\lim_{\alpha \to +\infty} g_{\max}(\alpha\mathbf{u}) = \lim_{\alpha \to +\infty} g_{i_{\mathbf{u}}}(\alpha\mathbf{u}) = +\infty \text{ for all } \mathbf{u} \neq \mathbf{0}$$

(c) We can find the minimal of $g_{\max}$ on $\{\mathbf{u} : \|\mathbf{u}\| = 1\}$, i.e.,

$$\min_{\|\mathbf{u}\|=1} g_{\max}(\mathbf{u}) = \epsilon > 0. \quad \text{(Weierstrass's Theorem)}$$

As the set is compact, there exists some $\mathbf{u}_0$ in $\{\mathbf{u} : \|\mathbf{u}\| = 1\}$ such that $g_{\max}(\mathbf{u}_0) = \epsilon$.

(d) If $g_{\max}(\mathbf{u}_1) \geq g_{\max}(\mathbf{u}_2)$, then

$$g_{\max}(\alpha\mathbf{u}_1) \geq g_{\max}(\alpha\mathbf{u}_2), \forall\, \alpha > 0.$$

To prove this claim, suppose $g_{\max}(\mathbf{u}_1) = g_{i_{\mathbf{u}_1}}(\mathbf{u}_1)$ and $g_{\max}(\mathbf{u}_2) = g_{j_{\mathbf{u}_2}}(\mathbf{u}_2)$. Since $\{([\mathbf{x}_i]_{1:r}, y_i)\}_{i=1}^n$ is not linearly separable, we have

$$\langle \mathbf{u}_1, \left[\overline{\mathbf{x}}_{i_{\mathbf{u}_1}}\right]_{1:r} \rangle < 0, \text{ if } i_{\mathbf{u}_1} \in \mathcal{L}^+;$$
$$\langle \mathbf{u}_1, \left[\overline{\mathbf{x}}_{i_{\mathbf{u}_1}}\right]_{1:r} \rangle > 0, \text{ if } i_{\mathbf{u}_1} \in \mathcal{L}^-.$$

Otherwise, $g_k(\mathbf{u}_i) \leq g_{i_{\mathbf{u}_1}}(\mathbf{u}_1) \leq \log 2$ for all $k \in [n]$, which leads to a contradiction that $\{([\mathbf{x}_i]_{1:r}, y_i)\}_{i=1}^n$ is not linearly separable. Similarly, we have

$$\langle \mathbf{u}_2, \left[\overline{\mathbf{x}}_{j_{\mathbf{u}_2}}\right]_{1:r} \rangle < 0, \text{ if } j_{\mathbf{u}_2} \in \mathcal{L}^+;$$
$$\langle \mathbf{u}_2, \left[\overline{\mathbf{x}}_{j_{\mathbf{u}_2}}\right]_{1:r} \rangle > 0, \text{ if } j_{\mathbf{u}_2} \in \mathcal{L}^-.$$

WLOG, suppose $y_i = y_j = 0$, then for any $\alpha > 1$ we have

$$g_{\max}(\alpha \mathbf{u}_1) - g_{\max}(\alpha \mathbf{u}_2) = \log \frac{1 + \exp(\alpha \langle \mathbf{u}_1, \left[\overline{\mathbf{x}}_{j_{\mathbf{u}_1}}\right]_{1:r} \rangle}{1 + \exp(\alpha \langle \mathbf{u}_2, \left[\overline{\mathbf{x}}_{j_{\mathbf{u}_2}}\right]_{1:r} \rangle} \geq \log 1 = 0.$$

That means

$$g_{\max}(\alpha \mathbf{u}_1) \geq g_{\max}(\alpha \mathbf{u}_2).$$

(e) Finally, we show that the solution to $\min_{\mathbf{u}} L^*(\mathbf{u})$ is in a compact ball. Consider $\mathbf{u}_0$ defined in (c). We have

$$\lim_{\alpha \to +\infty} g_{\max}(\alpha \mathbf{u}_0) = +\infty.$$

Thus, $\exists\, \alpha_0 > 0$ s.t.

$$g_{\max}(\alpha \mathbf{u}_0) \geq L^*(\mathbf{0}), \forall\, \alpha \geq \alpha_0.$$

Define $B_{\alpha_0}(\mathbf{0}) = \{\mathbf{u} : \|\mathbf{u}\| \leq \alpha_0\}$, then $\forall \mathbf{u} \notin B_{\alpha_0}(\mathbf{0})$, we have

$$L^*(\mathbf{u}) > g_{\max}(\mathbf{u}) \overset{(b)}{\geq} g_{\max}(\alpha_0 \frac{\mathbf{u}}{\|\mathbf{u}\|}) \overset{(d)}{\geq} g_{\max}(\alpha_0 \mathbf{u}_0) \geq L^*(\mathbf{0})$$

Hence the solution must be in a compact set $B_{\alpha_0}(\mathbf{0})$. Thus $L^*(\mathbf{u})$ attains its minimum at $\mathbf{u}^*$.

Let $(\mathbf{w}^*) = \begin{pmatrix} \mathbf{u}^* \\ \mathbf{0} \end{pmatrix}$. Next, we show that $L(\mathbf{w}^*) = \min_{\mathbf{w}} L(\mathbf{w})$. We have

$$\begin{aligned} L(\mathbf{w}) &= L^*((\mathbf{I}_r \quad \mathbf{M})\mathbf{w}) \\ &\geq L^*(\mathbf{u}^*) \\ &= L(\mathbf{w}^*). \end{aligned}$$

Therefore, $L(\mathbf{w})$ attains its minimum at $\mathbf{w}^*$.

Indeed, the minimum of $L(\mathbf{w})$ is not unique. If $\mathbf{M} = \mathbf{0}$, then $(\mathbf{w}^*) = \begin{pmatrix} \mathbf{u}^* \\ \mathbf{1} \end{pmatrix}$ is also an optimal solution.

It follows from the exercise 1.2 that the solution is unique if $\mathbf{rank}(\overline{\mathbf{X}}) = d + 1$.

■

Exercise 2: Programming Exercise: Naive Bayes 30pts

We provide you with a data set that contains spam and non-spam emails ("hw3_nb.zip"). Please use the Naive Bayes Classifier to detect the spam emails. Finish the following exercises by programming. You can use your favorite programming language.

1. Remove all the tokens that contain non-alphabetic characters.
2. Train the Naive Bayes Classifier on the training set according to Algorithm 1.
3. Test the Naive Bayes Classifier on the test set according to Algorithm 2.
4. Compute the confusion matrix, precision, recall and F1 score and then write down them in this file.

Algorithm 1 Training Naive Bayes Classifier

Input: The training set with the labels $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$.

$\mathcal{V} \leftarrow$ the set of distinct words and other tokens found in $\mathcal{D}$
for each target value c in the lables set $\mathcal{C}$ **do**
  $\mathcal{V}_c \leftarrow$ the training samples whose labels are c
  $P(c) \leftarrow \frac{|\mathcal{V}_c|}{|\mathcal{V}|}$
  $T_c \leftarrow$ a single document by concaatenating all training samples in $\mathcal{V}_c$
  $n_c \leftarrow |T_c|$
  for each word w_k in the vocabulary $\mathcal{V}$ **do**
    $n_{c,k} \leftarrow$ the number of times the word w_k occurs in T_c
    $P(w_k|c) = \frac{n_{c,k}+1}{n_c+|\mathcal{V}|}$
  end for
end for

Algorithm 2 Testing Naive Bayes Classifier

Input: An email $\mathbf{x}$. Let x_i be the i^{th} token in $\mathbf{x}$. $\mathcal{I} = \emptyset$.

for $i = 1, \dots, |\mathbf{x}|$ **do**
  if $\exists\, w_{ki} \in \mathcal{V}$ such that $w_{ki} = x_i$ **then**
    $\mathcal{I} \leftarrow \mathcal{I} \cup k_i$
  end if
end for
predict the label of $\mathbf{x}$ by

$$\hat{y} = \arg\max_{c \in \mathcal{C}} P(c) \prod_{i \in \mathcal{I}} P(w_{ki}|c)$$

Solution: "nb.py" is a simple referral code. The confusion matrix is $\begin{bmatrix} TP = 48 & FP = 1 \\ FN = 1 & TN = 241 \end{bmatrix}$.

The precision is

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{241}{242}.$$

The recall is

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{241}{242}.$$

The F1 score is

$$\text{F} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{241}{242}.$$

■

Exercise 3: Programming Exercise: Logistic Regression 30pts

We provide you with a data set that contains images fall into two classes ("hw3_lr.zip"). Please use the Logistic Regression to classify them. Finish the following exercises by programming. You can use your favorite programming language.

1. Choose a proper normalization method to process the data matrix.
2. Train the Logistic Regression Classifier on the training set.
3. Run the Logistic Regression Classifier on the test set. Please report the confusion matrix, precision, recall and F1 score in this file.

Solution: "lr.py" is a simple referral code. The confusion matrix is $\begin{bmatrix} TP = 1024 & FP = 9 \\ FN = 8 & TN = 1126 \end{bmatrix}$. The precision is

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{1024}{1033}.$$

The recall is

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{128}{129}.$$

The F1 score is

$$\text{F} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2048}{2065}.$$

■