Problem Set 2 for Machine Learning 15 Fall

Jingyuan Liu
AndrewId: jingyual
jingyual@andrew.cmu.edu

1 Bayes Optimal Classification

1.1 Determine the Bayes optimal classifier

Our goal is to minimize the risk, so classifier should choose the condition that have smaller risk:

$$f(x) = \begin{cases} 1 & if \quad \alpha p(f(x)=1, y=0) < \beta p(f(x)=0, y=1) \\ 0 & if \quad \alpha p(f(x)=1, y=0) > \beta p(f(x)=0, y=1) \end{cases}$$

1.2 Show the α and β

Using bayes rule, we can derive:

$$R = p(f(x)=1 \mid y=0) + p(f(x)=0 \mid y=1) \tag{1}$$

$$= \frac{p(f(x)=1, y=0)}{p(y=0)} + \frac{p(f(x)=0, y=1)}{p(y=1)} \tag{2}$$

Therefore, we can choose:

$$\alpha = \frac{1}{p(y=0)}, \beta = \frac{1}{p(y=1)} \tag{3}$$

1.3 Classification Problem

From the question, we would know:

$$p(x) = \begin{cases} 1-p & if \quad y=1 \quad and \quad x=0 \\ p & if \quad y=1 \quad and \quad x=1 \\ 1-q & if \quad y=0 \quad and \quad x=0 \\ q & if \quad y=0 \quad and \quad x=1 \end{cases}$$

$$p(y=0) = p(y=1) = \frac{1}{2} \tag{4}$$

When x = 0, we should choose y = 0, because 1 - p $<$ 1 - q. When x = 1, y =1:

$$f(x) = x, \qquad R = p(f(x)=1, y=0) + p(f(x)=0, y=1) = \frac{1}{2}(1-p+q) \tag{5}$$

2 Regularized Linear Regression Using Lasso

2.1 Show the J(w)

The goal is to find the w that minimize the error:

$$w^{\star} = \underset{w}{\operatorname{argmin}} \frac{1}{2} \|y - Xw\|^2 + \lambda \|w\|_1 \tag{6}$$

Therefore, to write in the form as required:

$$J_{\lambda}(w) = \frac{1}{2} \|y - Xw\|^2 + \lambda \|w\|_1 \tag{7}$$

$$= \frac{1}{2} \sum_{k}^{n} (y_k - \sum_{i}^{d} w_i X_{ki})^2 + \lambda \sum_{i} |w_i| \tag{8}$$

$$= \frac{1}{2} \sum_{k}^{n} (y_k^2 - 2y_k \sum_{i}^{d} w_i X_{ki} + (\sum_{i}^{d} w_i X_{ki})^2) + \lambda \sum_{i}^{d} |w_i| \tag{9}$$

As it notices that $X^T X = I$, so:

$$J_{\lambda}(W) = \frac{1}{2} \sum_{k}^{n} (y_k^2 - 2 \sum_{i}^{d} y_k w_i X_{ki} + \sum_{i}^{d} w_i^2) + \lambda \sum_{i}^{d} |w_i| \tag{10}$$

So to transfer to the form, we have:

$$g(y) = \frac{1}{2} y^2, \qquad f(X_{.i}, y, w_i, \lambda) = \frac{1}{2}((w_i^2 - 2yX_{.i}w_i) + \lambda |w_i| \tag{11}$$

2.2 Find $w_i^{\star}$ when $w_i^{\star} > 0$

As the previous function shows, and $w_i > 0$, we can have:

$$\frac{\partial J(W)}{\partial w_i} = w_i - (yX_i - \lambda) \tag{12}$$

The $w_i^{\star}$ is the best w_i that would make the J (w) smallest:

$$w_i^{\star} = \begin{cases} yX_i - \lambda & if \quad yX_i > \lambda \\ 0 & if \quad yX_i < \lambda \end{cases}$$

If $yX_i < \lambda$, then the minimal point that the gradient is zero could not be reached by the qudartic curve. Under this condition, the smaller the w_i, the smaller the J(w), so w_i = 0.

2.3 Find $w_i^{\star}$ when $w_i^{\star} < 0$

Similarily, we can get $w_i^{\star}$ under this condition:

$$\frac{\partial J(W)}{\partial w_i} = w_i - (yX_i + \lambda) \tag{13}$$

$$w_i^{\star} = \begin{cases} yX_i + \lambda & if \quad yX_i < -\lambda \\ 0 & if \quad yX_i > -\lambda \end{cases}$$

2.4 Find Condition $w_i^\star$ = 0

To conlude, under the two conditions, if we want the $w_i^\star$ be zero, we would need:

$$\lambda = \begin{cases} \lambda < -yX_i & if \quad yX_i < 0 \quad and \quad w_i <= 0 \\ \lambda > yX_i & if \quad yX_i > 0 \quad and \quad w_i >= 0 \end{cases}$$

$$\lambda > |yX_i| \tag{14}$$

We could find that this is really very reasonable answer, because adding lasso is to get a sparse parameter vector as mentioned. We could find that given a certain λ, if the absolute product of feature i and y yX_i is smaller than λ would be 0 to achieve the minimal likelihood J(W). With the lasso, those weights of features with "small" would temp to go to zero in the training.

2.5 Ridge Regression

As mentioned, we can have:

$$J_\lambda(W) = \frac{1}{2}\sum_k^n (y_k^2 - 2\sum_i^d y_k w_i X_{ki} + \sum_i^d w_i^2) + \lambda \sum_i^d \|w_i\|^2 \tag{15}$$

$$\frac{\partial J(W)}{\partial w_i} = (1+\lambda)w_i - yX_i \tag{16}$$

So if we want $w_i = 0$, then we need $yX_i = 0$. This is different from condition 4. Because the value of $w_i = 0$ only deepends y and X_i.

3 Multinomial Logistic Regression

3.1 Show the special form, logistic regression

Suppose the C = 2, then we have:

$$p(y = c^0 \mid x, W) = \frac{exp(w_{c0}^0 + w_c^{0T}x)}{exp(w_{c0}^0 + w_c^{0T}x) + exp(w_{c0}^1 + w_c^{1T}x)} \tag{17}$$

$$p(y = c^1 \mid x, W) = \frac{exp(w_{c0}^1 + w_c^{1T}x)}{exp(w_{c0}^0 + w_c^{0T}x) + exp(w_{c0}^1 + w_c^{1T}x)} \tag{18}$$

and we could transfer to:

$$p(y = c^0 \mid x, W) = \frac{1}{1 + exp(w_{c0}^1 - w_{c0}^0 + w_c^{1T}x - w_c^{0T}x)} \tag{19}$$

$$p(y = c^0 \mid x, W) = \frac{exp(w_{c0}^1 - w_{c0}^0 + w_c^{1T}x - w_c^{0T}x)}{1 + exp(w_{c0}^1 - w_{c0}^0 + w_c^{1T}x - w_c^{0T}x)} \tag{20}$$

We could see that multiclass Logistic regression reduce to logsitic regression when C = 2.

3.2 Multinomial Logistic Regression

Log Likelihood Function

We could derive the log likelihood based on the given form:

$$l(W) = log(\prod_i p(y_i \mid x_i, W)) \tag{21}$$

$$= log(\prod_i \prod_c p(y_i = c \mid x_i, W)) \tag{22}$$

Here we could use a denotation function, t_{ic}:

$$t_{ic} = \begin{cases} 1 & if \quad y_i == c \\ 0 & if \quad y_i! = c \end{cases}$$

With this denotation function:

$$l(w) = log(\prod_i \prod_c p(y_i = c \mid x_i, W)^{t_{ic}}) \tag{23}$$

$$= \sum_i \sum_c t_{ic} log(p(y_i = c \mid x_i, W)) \tag{24}$$

$$= \sum_i \sum_c t_{ic} log(\frac{exp(w_{c0} + w_c^T x)}{\sum_{c'} exp(w_{c'0} + w_{c'}^T x)}) \tag{25}$$

Derive the Gradient

To maximize the likelihood function, we need to derive the gradients for each weight:

$$g_c(W) = \frac{\partial l(W)}{\partial w_c} \tag{26}$$

$$= \frac{\partial \sum_i \sum_c t_{ic} log(p(y_i = c \mid x_i, W))}{\partial w_c} \tag{27}$$

$$= \sum_{i \in (y_i = c)} x_i - \sum_i x_i \frac{exp(w_c^T x_i)}{\sum_{c'} exp(w_{c'}^T x_i)}) \tag{28}$$

Derive the Hessian

To derive the hessian, we could do it based on the gradient:

$$H_{c,c'}(W) = \frac{\partial^2 l(W)}{\partial w_c \partial w_{c'}} \tag{29}$$

$$= \frac{\partial g_c(W)}{\partial w_{c'}} \tag{30}$$

$$= \frac{\partial \sum_i x_i (1 - p(y_i = c \mid x_i, W))}{\partial w_{c'}} \tag{31}$$

$$= \begin{cases} \sum_i x_i^2 p(y_i = c \mid x_i, W)(p(y_i = c' \mid x_i, W) - 1) & if \quad c == c' \\ \sum_i x_i^2 p(y_i = c \mid x_i, W) p(y_i = c' \mid x_i, W) & if \quad c! = c' \end{cases}$$

4 Perceptron Mistake Bounds

4.1 Show that $\langle w^t, w \rangle >= t\gamma$

$$\left\langle w^t, w \right\rangle = \left\langle w_{t-1} + y^t x^t, w \right\rangle \tag{32}$$

$$= \left\langle w^{t-1}, w \right\rangle + \left\langle y^t x^t, w \right\rangle \tag{33}$$

$$>= \left\langle w^{t-1}, w \right\rangle + \gamma \tag{34}$$

We can continously derive $\left\langle w^{t-1}, w \right\rangle$ to w^0, and there will be total t items. Therefore, we could

$$\left\langle w^t, w \right\rangle > t\gamma \tag{35}$$

4.2 Show that $\|w^t\|_2^2 <= tM^2$

$$\left\|w^t\right\|_2^2 = \left\langle w^t, w^t \right\rangle = \left\langle w^{t-1} + y^t x^t, w^{t-1} + y^t x^t \right\rangle \tag{36}$$

$$= \left\langle w^{t-1}, w^{t-1} \right\rangle + 2\left\langle w^{t-1}, y^t x^t \right\rangle + \left\langle y^t x^t, y^t x^t \right\rangle \tag{37}$$

We know that y^t, x^t is the misclassified cases, therefore $\left\langle w^{t-1}, y^t x^t \right\rangle < 0$. And y can only be 1 or -1. So we know that $\langle y^t x^t, y^t x^t \rangle$ is $\|x\|_2^2 < M^2$. So, we have:

$$\left\|w^t\right\|_2^2 <= \left\langle w^{t-1}, w^{t-1} \right\rangle + M^2 \tag{38}$$

We can derive w^t to w^0, so we have total t items, so we can prove that:

$$\left\|w^t\right\|_2^2 <= tM^2 \tag{39}$$

4.3 Prove the upper bound

We know that the meaning of $\langle a, b \rangle$ is related the $cos(\theta)$, which is the cos value of the angle between the vector a and b. So we have:

$$cos(\theta) = \frac{\langle w^t, w \rangle}{\|w_t\| \, \|w\|} \tag{40}$$

We know that $\|w\| = 1$, and take the form experssion into it, we have

$$cos(\theta) = \frac{t\gamma}{\sqrt{t}M} <= 1 \tag{41}$$

So we have that:

$$t <= \frac{M^2}{\gamma^2} \tag{42}$$

4.4 True or False

I think it is false. There should be a lot of classifiers that achieve zero error. But only $w = w^t$ and $t = \frac{M^2}{\gamma^2}$ will the classifier have margin γ.

5 Logistic Regression for Image Classification

5.1 Exploring the data

Run the modified code, and look at the variables stored in the memory

size of image

The size of image is 784 X 8 Byte.

range of labels

1 to 10

range of pixel values

0 to 1

max and min l2-norm

max: 17.1790, min: 3.5698

sparsity

we could find that 80.88% nodes are 0 value, so the data is sparse.

uniform

The max is 6742, the min is 5421. So it is uniformed.

5.2 Binary Logistic Regression

Without Regularization

The final objective function value is -897.275,

the $\|w\|^2$ is 19.08,

the trainning accuracy is 0.978551

the testing accuracy is 0.968246

trainning iterations is 674

With Regularization

The final objective function value is -1008.07,

the $\|w\|^2$ is 9.75082,

the trainning accuracy is 0.977466

the testing accuracy is 0.969254

trainning iterations is 326

Conclustion

We could find that adding regularization will lead the iteration faster to conrvege, get higher test performances, and have smaller norm of w.

5.3 Multiclass logistic regression

The final objective function value is -15269.5,

the $\|w\|^2$ is 47.7107,

the trainning accuracy is 0.931033

the testing accuracy is 0.925300

trainning iterations is 662

The visualization is as follows:

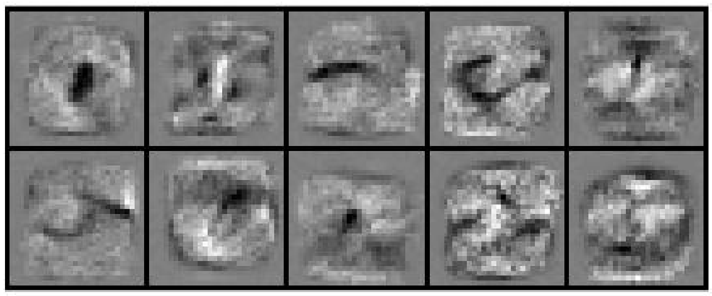

Figure 1: visualization

6 Collaboration

I dicussed with Zheng Chen with problem 2 on understanding finding the minimal for a quardratic function. And discussed with him on question 4 about using the cos value. And double checked the question 5 implementation.