Homework 1

Kai Qu
kq4ff

September 24, 2019

1 Let E be the event of the classifier making an error.

$$\begin{aligned}
P(Y=0|X=0) &= \frac{P(Y=1)P(X=0|Y=1)}{\sum_{y\in\{0,1\}} P(X=0,Y=y)} \\
&= (1-\beta)/(1-\beta+\alpha) \\
P(Y=0|X=1) &= \frac{P(Y=0)P(X=1|Y=0)}{\sum_{y\in\{0,1\}} P(X=1,Y=y)} \\
&= (1-\alpha)/(1-\alpha+\beta) \\
P(E) &= P(Y=0|X=1) + P(Y=1|X=0) \\
&= (1-\beta)/(1-\beta+\alpha) + (1-\alpha)/(1-\alpha+\beta)
\end{aligned}$$

2

Proof. Let D_1, D_2 be the two datasets, and $D_1 \cup D_2$ be the union of the two datasets (keep two copies of the sample even if both datasets contain the identical sample). Let $L_1(\theta), \theta_1, L_2(\theta), \theta_2$ be the negative log-likelihood functions and unregularized logistic regression coefficients for D_1 and D_2 respectively. Let $L(\theta), \theta^*$ be the negative log-likelihood function and the unregularized logistic regression coefficients for $D_1 \cup D_2$. Let θ_j be a particular feature.

Since $D_1 \cup D_2$ is the union of D_1 and D_2, we have

$$L(\theta) = L_1(\theta) + L_2(\theta)$$
$$\frac{\partial}{\partial \theta_j} L(\theta) = \frac{\partial}{\partial \theta_j} L_1(\theta) + \frac{\partial}{\partial \theta_j} L_2(\theta)$$

When $\theta = \theta_1$, $\frac{\partial}{\partial \theta_j} L_1(\theta) = 0$

$$\frac{\partial}{\partial \theta_j} L(\theta_1) = \frac{\partial}{\partial \theta_j} L_2(\theta_1)$$

When $\theta = \theta_2$, $\frac{\partial}{\partial \theta_j} L_2(\theta) = 0$

$$\frac{\partial}{\partial \theta_j} L(\theta_2) = \frac{\partial}{\partial \theta_j} L_1(\theta_2)$$

Case 1: When $\theta_j^{(1)} < \theta_j^{(2)}$:

Since the loss function is convex, $\frac{\partial}{\partial \theta_j} L(\theta_2) = \frac{\partial}{\partial \theta_j} L_2(\theta_2)$ achieves minimum at θ_2. Because θ_1 is at the left of θ_2,

$$\frac{\partial}{\partial \theta_j} L(\theta_2) = \frac{\partial}{\partial \theta_j} L_1(\theta_2) > 0$$

Similarly, $\frac{\partial}{\partial \theta_j} L(\theta_1) = \frac{\partial}{\partial \theta_j} L_1(\theta_1)$ achieves the minimum at θ_1. Because θ_2 is at the right of θ_1,

$$\frac{\partial}{\partial \theta_j} L(\theta_1) = \frac{\partial}{\partial \theta_j} L_2(\theta_1) < 0$$

Since $\frac{\partial}{\partial \theta_j} L(\theta)$ is the partial derivative of $L(\theta)$, which is convex, $\frac{\partial}{\partial \theta_j} L(\theta)$ is continuous. Moreover, we have $\theta_j^2 > \theta_j^1$ and $\frac{\partial}{\partial \theta_j} L(\theta_1) = \frac{\partial}{\partial \theta_j} L_2(\theta_1) < 0$ and $\frac{\partial}{\partial \theta_j} L(\theta_2) = \frac{\partial}{\partial \theta_j} L_1(\theta_2) > 0$. By Intermediate Value Theorem, there exists $\theta_j^1 \leq \theta_j^* \leq \theta_j^2$ such that $\frac{\partial}{\partial \theta_j} L(\theta) = 0$. Since L is convex, θ_1^* is the solution for feature j. Also, $min(\theta_j^1, \theta_j^2) \leq \theta_j^* \leq max(\theta_j^1, \theta_j^2)$.

Case 2: When $\theta_j^{(1)} > \theta_j^{(2)}$:

Since the loss function is convex, $\frac{\partial}{\partial \theta_j} L(\theta_1) = \frac{\partial}{\partial \theta_j} L_2(\theta_1)$ achieves minimum at θ_2. Because θ_1 is at the right of θ_2,

$$\frac{\partial}{\partial \theta_j} L(\theta_1) = \frac{\partial}{\partial \theta_j} L_2(\theta_1) > 0$$

Similarly, $\frac{\partial}{\partial \theta_j} L(\theta_2) = \frac{\partial}{\partial \theta_j} L_1(\theta_2)$ achieves the minimum at θ_1. Because θ_1 is at the right of θ_2,

$$\frac{\partial}{\partial \theta_j} L(\theta_2) = \frac{\partial}{\partial \theta_j} L_1(\theta_2) < 0$$

Since $\frac{\partial}{\partial \theta_j} L(\theta)$ is the partial derivative of $L(\theta)$, which is convex, $\frac{\partial}{\partial \theta_j} L(\theta)$ is continuous. Moreover, we have $\theta_j^1 > \theta_j^2$ and $\frac{\partial}{\partial \theta_j} L(\theta_1) = \frac{\partial}{\partial \theta_j} L_2(\theta_1) > 0$ and $\frac{\partial}{\partial \theta_j} L(\theta_2) = \frac{\partial}{\partial \theta_j} L_1(\theta_2) < 0$. By Intermediate Value Theorem, there exists $\theta_j^2 \leq \theta_j^* \leq \theta_j^1$ such that $\frac{\partial}{\partial \theta_j} L(\theta) = 0$. Since L is convex, θ_1^* is the solution for feature j. Also, $min(\theta_j^1, \theta_j^2) \leq \theta_j^* \leq max(\theta_j^1, \theta_j^2)$.

Therefore, by the two cases above, we can conclude that $min(\theta_j^1, \theta_j^2) \leq \theta_j^* \leq max(\theta_j^1, \theta_j^2)$. $\square$

3

Proof. Let $\hat{\theta} = \arg\min_\theta L(\theta)$, where L is the negative log likelihood function. Let $\theta^* = \arg\min_\theta L(\theta) + \lambda||\theta||_2^2, \lambda > 0$. Given the conditions above, we have:

$$
\begin{aligned}
L(\theta^*) + \lambda||\theta^*||_2^2 &\leq L(\hat{\theta}) + \lambda||\hat{\theta}||_2^2 \\
L(\hat{\theta}) &\leq L(\theta^*) \\
L(\hat{\theta}) + \lambda||\theta^*||_2^2 &\leq L(\hat{\theta}) + \lambda||\hat{\theta}||_2^2 \\
\lambda||\theta^*||_2^2 &\leq \lambda||\hat{\theta}||_2^2 \\
||\theta^*||_2^2 &\leq ||\hat{\theta}||_2^2
\end{aligned}
$$

Therefore, the conclusion $||\theta^*||_2^2 \leq ||\hat{\theta}||_2^2$ is proved. □

4 Let y be the ground truth values and $\hat{y}$ be the predicted values. First, we prove that the F-measure $F(y, \hat{y}) = 2pr/(p + r)$ is less than or equal to $(p + r)/2$, where p and r are the precision and the recall of the logistic regression. Assume that the count of the True Positive (TP) is greater than 0. Otherwise, $p = r = 0$ and F-measure $F(y, \hat{y}) = 2pr/(p+r)$ will not be defined.

Proof. When $p > 0, r > 0$, in order to prove $2pr/(p + r) \leq (p + r)/2$, we can prove its equivalence $(p + r)^2 - 4pr \geq 0$

$$
\begin{aligned}
(p + r)^2 - 4pr &= p^2 + 2pr + r^2 - 4pr \\
&= p^2 - 2pr + r^2 \\
&= (p - r)^2 \\
&\geq 0
\end{aligned}
$$

Therefore, since $2pr/(p + r) \leq (p + r)/2$ for all possible values of p and r, we conclude that $2pr/(p + r) \leq (p + r)/2$. □

Next, we try to prove that $2pr/(p + r) = (p + r)/2$ if and only if $p = r$.

Proof. $\Rightarrow$ if $2pr/(p + r) = (p + r)/2$,

$$
\begin{aligned}
2pr/(p + r) &= (p + r)/2 \\
(p + r)^2 - 4pr &= 0 \\
(p - r)^2 &= 0 \\
p &= r
\end{aligned}
$$

$\Leftarrow$ if $p = r$,

$$\begin{aligned} 2pr/(p+r) &= 2 * r * r/2r \\ &= r \\ &= (r+r)/2 \\ &= (p+r)/2 \end{aligned}$$

Thus, we conclude that $2pr/(p + r) = (p + r)/2$ if and only if $p = r$. $\square$

5 I tried three different sets of parameters for $CountVectorizer$ and $Logistic\,regression$. The parameters and results are as below:

1. $vectorizer = CountVectorizer(lowercase = True, min\_df = 7e{-}5, ngram\_range = (1, 2), max\_features = 10000)$
$classifier = LogisticRegression(solver = "saga", multi\_class = "multinomial", penalty = "l2")$
Training Set Accuracy: 0.811650
Development Set Accuracy: 0.6502

2. $vectorizer = CountVectorizer(lowercase = True, min\_df = 7e - 5)$
$classifier = LogisticRegression(solver = "sag", multi\_class = "multinomial", penalty = "l2")$
Training Set Accuracy: 0.767
Development Set Accuracy: 0.6396

3. $vectorizer = CountVectorizer(lowercase = True, min\_df = 7e - 5)$
$classifier = LogisticRegression(solver = "saga", multi\_class = "multinomial", penalty = "l2")$
Training Set Accuracy: 0.7342
Development Set Accuracy: 0.642

The first set of parameters gave the best result regarding to the Development Set Accuracy.