CS181 Homework 2

RUITAO (TOBY) DU

02/06/2015

1 Alternative Softmax Parameterization

In *softmax* function,

$$\Pr(t_k = 1 \mid \boldsymbol{x}, \{\boldsymbol{w}_{k'}\}_{k'=1}^{K}) = \frac{\exp\{\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x}\}}{\sum_{k'=1}^{K}\exp\{\boldsymbol{w}_{k'}^\mathsf{T}\boldsymbol{x}\}}\,.$$

We have a set of K vectors. However, it is obvious that no matter how we assign $\boldsymbol{w}_k (k = 1, 2, ..., K)$, the sum of these probability always equal to 1.

$$\sum_{k=1}^{K}\Pr(t_k = 1 \mid \boldsymbol{x}, \{\boldsymbol{w}_{k'}\}_{k'=1}^{K}) = \sum_{k=1}^{K}\frac{\exp\{\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x}\}}{\sum_{k'=1}^{K}\exp\{\boldsymbol{w}_{k'}^\mathsf{T}\boldsymbol{x}\}} \equiv 1\,.$$

Hence we can loose the constraint and use a set of K-1 vectors to give the same distribution. Without loss of generality, I devide numerator and denominator by $\exp\{\boldsymbol{w}_K^\mathsf{T}\boldsymbol{x}\}$ I assign $\boldsymbol{w}'_k = \boldsymbol{w}_k - \boldsymbol{w}_K$ and assign 0 to $\boldsymbol{w}'_K$. so we can have our alternative *softmax* function:

$$\Pr(t_k = 1 \mid \boldsymbol{x}, \{\boldsymbol{w}'_{k'}\}_{k'=1}^{K-1}) = \begin{cases} \dfrac{\exp\{\boldsymbol{w}_k'^\mathsf{T}\boldsymbol{x}\}}{1+\sum_{k'=1}^{K-1}\exp\{\boldsymbol{w}_{k'}'^\mathsf{T}\boldsymbol{x}\}} & k < K \\ \dfrac{1}{1+\sum_{k'=1}^{K-1}\exp\{\boldsymbol{w}_{k'}'^\mathsf{T}\boldsymbol{x}\}} & k = K \end{cases}.$$

2 Maximum Likelihood Training Objective

To help me write the likelihood function, I build following vector $\boldsymbol{P}$ because t_n is one-hot label. I can use the dot product of these two vectors to find the probability

$$\boldsymbol{P}_n = \begin{bmatrix} \Pr(t_1 = 1 \mid \boldsymbol{x}_n, \{\boldsymbol{w}_{k'}\}_{k'=1}^{K-1}) \\ \Pr(t_2 = 1 \mid \boldsymbol{x}_n, \{\boldsymbol{w}_{k'}\}_{k'=1}^{K-1}) \\ \vdots \\ \Pr(t_K = 1 \mid \boldsymbol{x}_n, \{\boldsymbol{w}_{k'}\}_{k'=1}^{K-1}) \end{bmatrix} = \begin{bmatrix} \frac{\exp\{\boldsymbol{w}_1^\top \boldsymbol{x}_n\}}{1+\sum_{k'=1}^{K-1}\exp\{\boldsymbol{w}_{k'}^\top \boldsymbol{x}_n\}} \\ \frac{\exp\{\boldsymbol{w}_2^\top \boldsymbol{x}_n\}}{1+\sum_{k'=1}^{K-1}\exp\{\boldsymbol{w}_{k'}^\top \boldsymbol{x}_n\}} \\ \vdots \\ \frac{\exp\{\boldsymbol{w}_{K-1}^\top \boldsymbol{x}_n\}}{1+\sum_{k'=1}^{K-1}\exp\{\boldsymbol{w}_{k'}^\top \boldsymbol{x}_n\}} \\ \frac{1}{1+\sum_{k'=1}^{K-1}\exp\{\boldsymbol{w}_{k'}^\top \boldsymbol{x}_n\}} \end{bmatrix}$$

The likelihood funtion would be

$$\begin{aligned} Likelihood &= Pr(D \mid \{\boldsymbol{w}_{k'}\}_{k'=1}^{K-1}) \\ &= \prod_{n=1}^{N} Pr(\boldsymbol{t}_n \mid \boldsymbol{x}_n, \{\boldsymbol{w}_{k'}\}_{k'=1}^{K-1}) \\ &= \prod_{n=1}^{N} \boldsymbol{P}_n \cdot \boldsymbol{t}_n \end{aligned}$$

where $\boldsymbol{t}_n$ is the one-hot label

3 Gradient of Alternative Softmax Parameterization

When $k < K$,

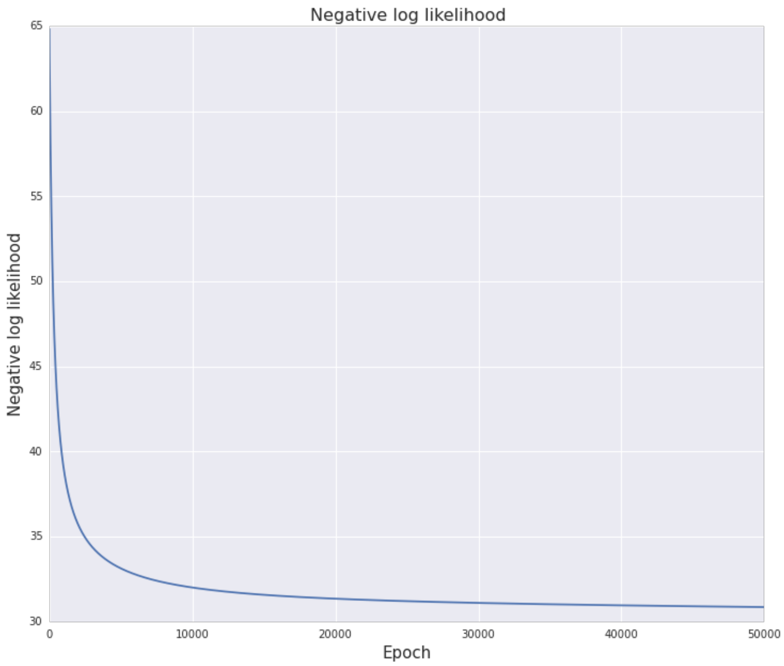
$$l(D \mid \{\boldsymbol{w}_{k'}\}_{k'=1}^{K-1}) = \sum_{n=1}^{N} \boldsymbol{P}_n \cdot \boldsymbol{t}_n$$

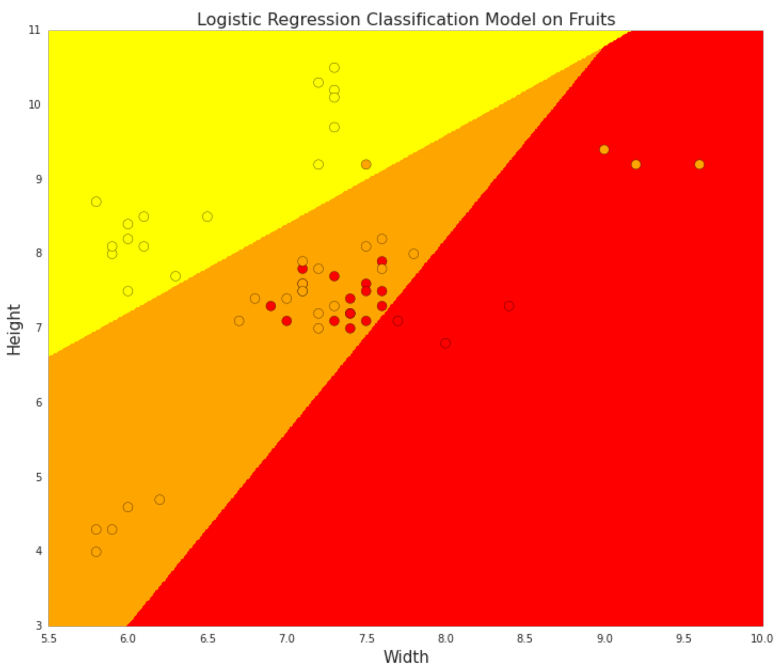
When we calcultate the gradient of $\boldsymbol{w}_k$, we can eliminate other vector $\boldsymbol{w}_i\,(i \neq k)$. k_n is the class of data $x_n, t_n\,(n = 1, 2, \ldots, N)$

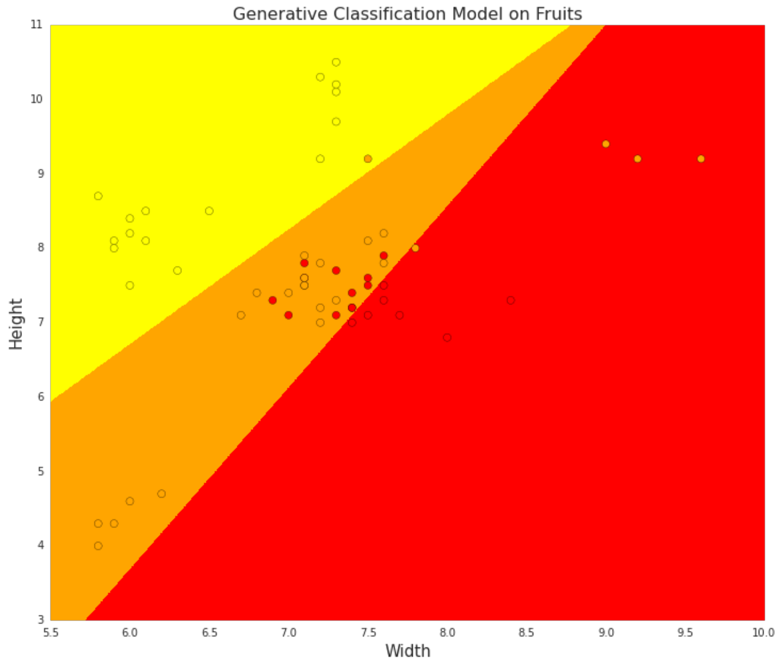
$$\begin{aligned}
\nabla_{\boldsymbol{w}_k} l(D \mid \{\boldsymbol{w}_{k'}\}_{k'=1}^{K-1}) &= \nabla_{\boldsymbol{w}_k} \sum_{n=1}^{N} \ln \boldsymbol{P}_n \cdot \boldsymbol{t}_n \\
&= \nabla_{\boldsymbol{w}_k} \sum_{n=1}^{N} \ln \frac{\exp\{\boldsymbol{w}_{k_n}^{\mathsf{T}} \boldsymbol{x}_n\}}{1 + \sum_{k'=1}^{K-1} \exp\{\boldsymbol{w}_{k'}^{\mathsf{T}} \boldsymbol{x}_n\}} \\
&= \nabla_{\boldsymbol{w}_k} \sum_{n=1}^{N} (\boldsymbol{w}_{k_n}^{\mathsf{T}} \boldsymbol{x}_n - \ln(1 + \sum_{k'=1}^{K-1} \exp\{\boldsymbol{w}_{k'}^{\mathsf{T}} \boldsymbol{x}_n\})) \\
&= \sum_{\boldsymbol{t}_n \in k} \boldsymbol{x}_n - \sum_{n=1}^{N} \nabla_{\boldsymbol{w}_k} \ln(1 + \sum_{k'=1}^{K-1} \exp\{\boldsymbol{w}_{k'}^{\mathsf{T}} \boldsymbol{x}_n\}) \\
&= \sum_{\boldsymbol{t}_n \in k} \boldsymbol{x}_n - \sum_{n=1}^{N} \frac{\exp\{\boldsymbol{w}_{k}^{\mathsf{T}} \boldsymbol{x}_n\} \boldsymbol{x}_n}{1 + \sum_{k'=1}^{K-1} \exp\{\boldsymbol{w}_{k'}^{\mathsf{T}} \boldsymbol{x}_n\}}
\end{aligned}$$

$\boldsymbol{t}_n \in k$ means $\boldsymbol{t}_{n,k} = 1$

4 Classifying Fruit







In the first model, I use gradient ascent on the log likelihood to solve w. In the latter model, I use share covariance matrix so as to have a linear dicision boundary, which is easy to compare between these two methods. From these plots, we know that the generative classification model is better and the decision boundary is more accurate. I think the reason is that the number of our data is not too much so when we use generative model, we assume the prior first. This prior is important when we don't have enough data.

5 Calibration

8 hours