CS229 Fall 2017
Problem Set #2: Supervised Learning II

Author: LFhase rimemosa@163.com

Logistic Regression: Training stability

(a) Training model on dataset A costs far more less time than that on dataset B, which means that training on dataset B does't converge.

(b) Let's plot the training results after 10000, 20000, 30000, 40000 iterations.

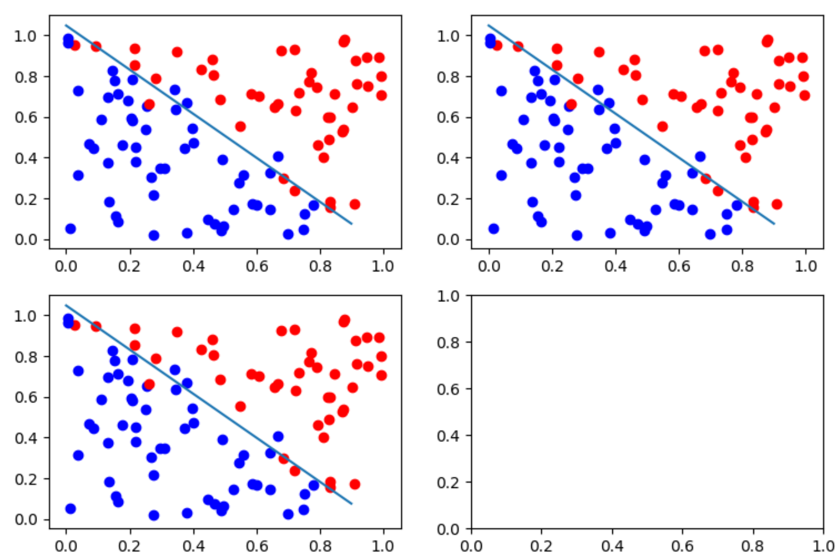

Figure 1: Training Results on Dataset A

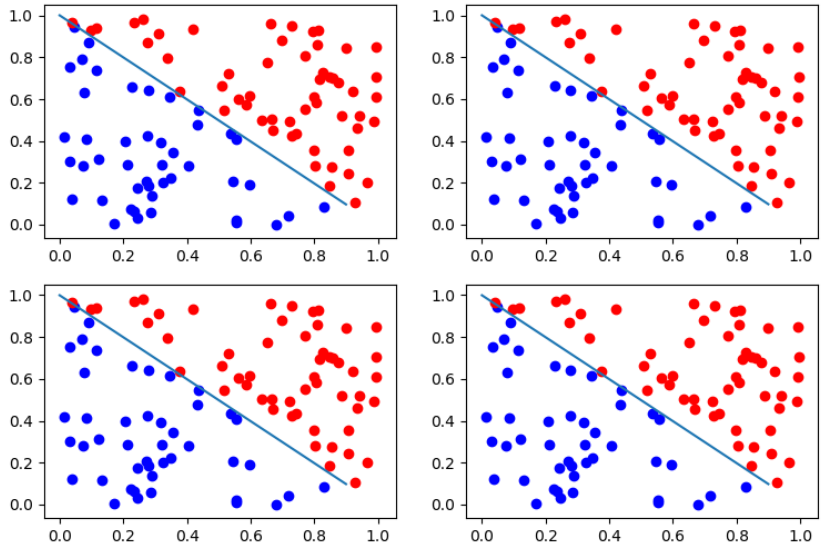

Figure 2: Training Results on Dataset B

From the above two figures, we can see that data on dataset B is hardly to separate (Bad Linearly Separability), which may be the main issue resluting nonconvergence.

(c) i No. Using a different learning rate only changes the learning speed here, but it won't change the fact that the algorithm has to find the hyperline in hardly separable data.

ii No. The same to the former.

iii Yes. It will stop $||\theta||$ being infinitely large.

iv No. It doesn't change the linearly separability.

v Yes. It will expand the feature space, which may let the data linearly separable.

(d) It's vulnerable. With hinge loss, using slack variables, the formulation will be changed into what's be induced in class.

Model Calibration

(a) Firstly, we write the log-likelihood function of Logistic Regression:

$$J(\theta) = \sum_{i=1}^{m}(y^{(i)}logh(x^{(i)}) + (1 - y^{(i)})log(1 - h(x^{(i)})))$$

Then, let

$$\frac{\partial J(\theta)}{\partial \theta} = 0$$

We get

$$\sum_{i=1}^{m}(y^{(i)} - h(x^{(i)}))x_j^{(i)} = 0$$

Because $x_0 = 1$ for all training examples, so $|X_{m\times n}| \neq 0$ and $y^{(i)} - h(x^{(i)}) = 0$, which means

$$\sum_{i=1}^{m} h(x^{(i)}) = \mathbf{1}\{y^{(i)} = 1\}$$

The property described in problem statement gets proved.

(b) Perfect calibration means the model has a good performance in training data, which doesn't ensure the model achieves perferct accuray in other conditions. However, once a model achieves perferct accuray, it should be perferctly calibrated.

(c) The new log-likelihood function will become

$$J'(\theta) = J(\theta) + c||\theta||^2$$

Let

$$\frac{\partial J'(\theta)}{\partial \theta} = 0$$

We get

$$\frac{\partial J(\theta)}{\partial \theta} + 2c\theta = 0$$

which means

$$\sum_{i=1}^{m}(y^{(i)} - h(x^{(i)}))x_j^{(i)} + 2c\theta_0 = 0$$

and

$$\sum_{i=1}^{m} h(x^{(i)}) = \mathbf{1}\{y^{(i)} = 1\} + 2c\theta_0$$

The left part in Model Calibration equation will get $2c\theta_0$ bias.

Bayesian Logistic Regression and weight decay

Assume that $||\theta_{MAP}||_2 > ||\theta_{ML}||_2$, we can get

$$p(\theta_{MAP}) < p(\theta_{ML})$$

thus

$$p(\theta_{MAP})\Pi_{i=1}^m p(y^{(i)}|x^{(i)};\theta_{MAP}) < p(\theta_{ML})\Pi_{i=1}^m p(y^{(i)}|x^{(i)};\theta_{MAP})$$

with

$$\Pi_{i=1}^m p(y^{(i)}|x^{(i)};\theta_{MAP}) < \Pi_{i=1}^m p(y^{(i)}|x^{(i)};\theta_{ML})$$

we get

$$p(\theta_{MAP})\Pi_{i=1}^m p(y^{(i)}|x^{(i)};\theta_{MAP}) < p(\theta_{ML})\Pi_{i=1}^m p(y^{(i)}|x^{(i)};\theta_{ML})$$

which is contradicted with the definition of θ_{MAP}.

Constructing kernels

According to Mercer's theorem, if $K(x,z)$ is a kernel, then we have $\mu^T M\mu \geq 0$.

(a) Yes. Since $K_1 \geq 0$ and $K_2 \geq 0$, so $K_1 + K_2 \geq 0$.

(b) No. $K_1 \geq 0$ and $K_2 \geq 0$ doesn't necessarily mean $K_1 - K_2 \geq 0$.

(c) If $a \geq 0$, the answer is 'Yes'.

(d) If $a \geq 0$, the answer is 'No'.

(e) Since K_1 and K_2 are valid kernels, we can let ϕ_1 and ϕ_2 be their feature map function and ϕ be K's. Then we have

$$\begin{aligned} K(x,z) &= \phi_1(x)^T\phi_1(z)\phi_2(x)^T\phi_2(z) \\ &= \sum_{i,j}\phi_1(x)_i\phi_1(z)_i\phi_2(x)_j\phi_2(z)_j \\ &= \sum_{i,j}[\phi_1(x)_i\phi_2(x)_j][\phi_2(x)_j\phi_1(z)_i] \end{aligned}$$

If let $\phi_{i,j} = \phi_{1i}\phi_{2j}$, then we can easily write the inner product formula of $K(x,z)$. Through the definition of Kernel, we know that $K(x,z)$ is a valid kernel.

(f) Yes, since $K(x,z) = (\sum_{i=1}\mu_i f(x^{(i)}))^2 \geq 0$.

(g) Yes, since $K(x,z) = \sum_{i,j}\mu_i\mu_j K_3 \geq 0$.

(h) Yes, since $K(x,z) = \sum_{i,j}\mu_i\mu_j(\sum_t a_t {K_1}^t) \geq 0$.

Kernelizing the Perceptron

(a) gg