CS229 Problem Set 2

Tianyu Du

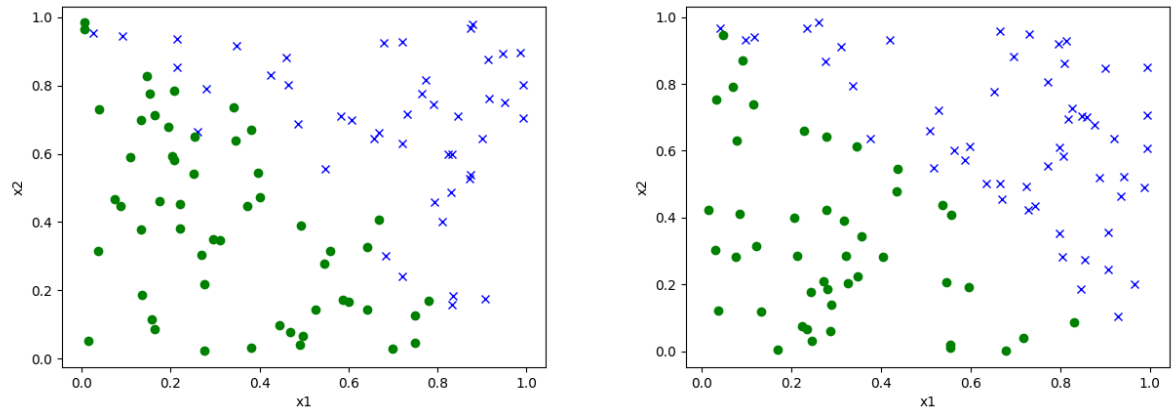
Tuesday 23rd July, 2019

1 Problem 1 Logistic Regression: Training Stability

1.1 1(a)

Comment The parameter trained on dataset A converges after around 30,000 iterations of gradient ascent. In contrast, on dataset B, the training algorithm does not reach convergence within a reasonable amount of time.

1.2 1(b)



(a) Dataset A (b) Dataset B

Figure 1: Two Datasets

Comment From the plot of two datasets we can see that dataset B is more *linearly separable* than dataset A. There exists intermixing of data points belonging to both classes in dataset A, but fewer can be observed in dataset B. The linear separability of dataset B causes the instability of logistic regression on it. In the scenario of dataset B, we claim that the gradient descent algorithm increases θ indefinitely without convergence.

Proof. Let $\mathcal{A}$ and $\mathcal{B}$ denote two datasets respectively. And the log-likelihood looks like

$$\ell(\theta) = \sum_{i=1}^{n} y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right) \tag{1.1}$$

$$= \underbrace{\sum_{i:y^{(i)}=0} \log\left(1 - h_\theta(x^{(i)})\right)}_{P} + \underbrace{\sum_{i:y^{(i)}=1} \log h_\theta(x^{(i)})}_{Q} \tag{1.2}$$

Suppose the threshold for classification is 0.5, and dataset $\mathcal{B}$ can be separated perfectly with some $\theta \in \mathbb{R}^d$. That's, for (almost) every sample $i \in \mathcal{B}$ such that $y^{(i)} = 0$, $\theta^T x^{(i)} < 0$ and for (almost) every $y^{(i)} = 1$, $\theta^T x^{(i)} > 0$. As a result, for every θ^t that separate dataset $\mathcal{B}$ almost perfectly, suppose θ^t is inflated to $\theta^{t+1} := (1+\varepsilon)\theta^t$ for some $\varepsilon > 0$, for those negative samples ($y^{(i)} = 0$), $\theta^T x^{(i)}$ changes to $(1+\varepsilon)\theta^T x^{(i)} < \theta^T x^{(i)} < 0$, and $h_\theta(x^{(i)})$ decreases. As a result, P increases. Similarly, for those positive samples, $\theta^T x^{(i)}$ changes to $(1+\varepsilon)\theta^T x^{(i)} > \theta^T x^{(i)} > 0$, and $h_\theta(x^{(i)})$, so does Q.

It is shown that for each θ^t separates dataset perfectly (or at least almost perfectly, because if the misclassified group is small, we can safely ignore the contribution to likelihood function from samples from the misclassified group), the log likelihood is increased when θ is inflated, and the

resulted θ^{t+1} is still a (almost) perfect-separating parameter. Also, there is no upper bound on $||\theta||$, therefore, the gradient ascend algorithm will run indefinitely to increase $\ell(\theta)$ by inflation the norm of θ.

For dataset like $\mathcal{A}$, for each θ, the misclassified group provides significant contributions to $\ell(\theta)$, inflating θ comes with the cost of reduced likelihood on samples from the misclassified groups. By the nature of log function, the cost of likelihood reduction on misclassified group will eventually overcome the the likelihood gain by expanding θ, and $\ell(\theta)$ cannot be risen to infinity on this kind of datasets. Therefore, the algorithm stops at some θ^*. ■

1.3 1(c)

Comment

(i) (NO) A different constant learning rate won't help because the $\nabla_\theta \ell(\theta)$ does not vanish, and the update step $\alpha \nabla_\theta \ell(\theta)$ is always significant. Therefore, convergence is still not achievable.

(ii) (NO) Annealing the learning rate will not mitigate the problem because this does not change the fact that the maximum of ℓ is achieved when $\theta \to \infty$.

(iii) (NO) Linear scaling does not change the fact that dataset is well (linearly) separable, θ is still going to explode indefinitely.

(iv) (YES) Adding regularization helps. As mentioned before, the main cause of interminability of gradient ascent is $\ell(\theta)$ can always be increased by inflating θ (i.e. increasing $||\theta||$). The regularizing term prevents θ from exploding and enforce the convergence.

(v) (YES) Give the variance of noise is large enough, the noise helps eliminate the linear separability of data, so that the dataset becomes intermixing.

1.4 1(d)

Comment SVM is more robust against linearly separable dataset. SVM is seeking to construct a hyperplane $w^T x + b$ to maximize the geometric margin γ, while controlling $||w|| = 1$. Note that changing b would shift the position of the hyperplane, therefore at optimal, $b^* < \infty$. Also, the performance measure of SVM using geometric margin is invariant to scales of (w, b). Therefore, given the constraint on $||w||$, SVM using geometric margin does not suffer from the parameter exploding problem (optimal is achieved when $\theta \to \infty$), and the optimal (w^*, b^*) will both be finite.

2 Problem 2 Spam Classification

2.1 2(a)

Result `Size of dictionary: 1722`

2.2 2(b)

Result `Naive Bayes had an accuracy of 0.978494623655914 on the testing set`

2.3 2(c)

Result `The top 5 indicative words for Naive Bayes are: ['claim', 'won', 'prize', 'tone', 'urgent!']`

2.4 2(d)

Result `The optimal SVM radius was 0.1`