Comp 540 Machine Learning

Yue He, Peng Yang

Feruary 2019

On my honor, I have neither given nor received any unauthorized aid on this assignment.

1 Intuitions about support vector machines

(a) A classifier with big margin makes no low certainty classification decisions. A large margin also corresponds to a regularization of SVM weights which prevents over-fitting. With large margin, our predictions are better on the test data by not over-fitting the model to the training data.

(b) No. Moving points which are not support vectors further away from the decision boundary will not effect the SVM's hinge loss. Since we know the hinge loss function is

$$\min_{\theta,\theta_0} \frac{1}{2}||\theta||^2 + C\sum_{i=1}^{m} max(0, 1 - y^{(i)}h_\theta(x^{(i)})$$

note that points which are not support vectors denotes $y^{(i)}h_\theta(x^{(i)} \geq 1 \Rightarrow max(0, 1 - y^{(i)}h_\theta(x^{(i)}) = 0$. Thus, when we move those points further away from the decision boundary, $max(0, 1 - y^{(i)}h_\theta(x^{(i)})$ still remain 0, so it will not effect the SVM's hinge loss.

2 Fitting an SVM classifier by hand

First, we know that

$$\phi(x^{(1)}) = (1, 0, 0), \quad y^{(1)} = -1$$
$$\phi(x^{(2)}) = (1, 2, 2), \quad y^{(1)} = 1$$

(a) Since we know that θ is perpendicular to the decision boundary which means θ is parallel to the vector $\vec{a} = (\phi(x^{(2)}), \phi(x^{(1)})) = \phi(x^{(2)}) - \phi(x^{(1)}) = (0, 2, 2) = c\theta$, where c is a constant. Let $c = \frac{1}{2}$, we get $\theta = (0, 1, 1)$.

(b) Since we know the only two points $\phi(x^{(1)})$ and $\phi(x^{(2)})$ are the support vectors, the value of the margin is actually the distance between those two points.

$$\begin{aligned} d &= d(\phi(x^{(1)}), \phi(x^{(2)})) \\ &= \sqrt{0^2 + 2^2 + 2^2} = 2\sqrt{2} \end{aligned}$$

(c) Based on the results above, we know $d = \frac{2}{||\theta||} = 2\sqrt{2} \Rightarrow ||\theta|| = \frac{\sqrt{2}}{2}$. Besides, we know that θ is actually perpendicular to $\vec{a}$, so we let $\theta = (0, a, a)^T$, where a is a constant.

$$\begin{aligned}\sqrt{(a^2 + a^2)} &= \frac{\sqrt{2}}{2} \\ \Rightarrow a &= \frac{1}{2} \\ \Rightarrow \theta &= (0, \frac{1}{2}, \frac{1}{2})^T\end{aligned}$$

(d) Based on the constrain,

$$\begin{aligned}&y^{(1)}(\theta^T \phi(x^{(1)}) + \theta_0) \geq 1 \\ &y^{(2)}(\theta^T \phi(x^{(2)}) + \theta_0) \geq 1 \\ \Rightarrow &-\theta_0 \geq 1 \\ &2 + \theta_0 \geq 1 \\ \Rightarrow &-1 \leq \theta_0 \leq -1 \\ \Rightarrow &\theta_0 = -1\end{aligned}$$

(e) Let Z denotes the decision boundary,

$$\begin{aligned}Z &= \theta^T \phi(x) + \theta_0 \\ &= \frac{\sqrt{2}}{2}x + \frac{1}{2}x^2 - 1\end{aligned}$$

3 Support vector machines for binary classification

3.3 Explain how you chose the parameters for training the SVM calculate each C, learning rate and iteration rate in for loop to find the parameter value with smallest loss and best accuracy. Best accuracy = 0.967; best C = 3; best learning rate = 0.00001; best iteration rate = 10000; best sigma = 10 Should X be scaled? Yes, to remove the dimensional. Should X be kernelized? Yes, we have test that without kernelized, the acuracy is low. And to make algorithm run faster, we choose to use kernel methods.

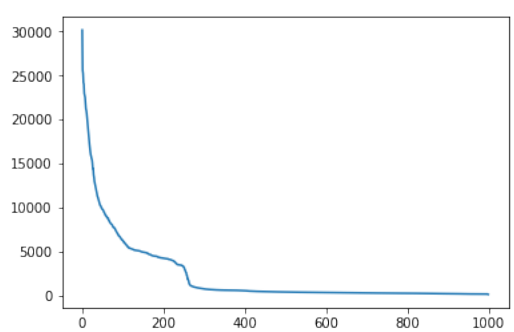

Figure 1: iteration number with loss

4 Support vector machines for multi-class classification

4A Since the loss function of SVM is not strictly differential, the gradient we get may not be accurate. So the difference between the result we get and numerical methods result is reasonable.

4E training time of SVM is 82.301949s. In Softmax, the time is around 14s. So SVM takes longer time. For accuracy, SVM is 0.385184 and Softmax is 0.412000. Softmax shows better than SVM in accuracy. The visualization results of two methods are similar.

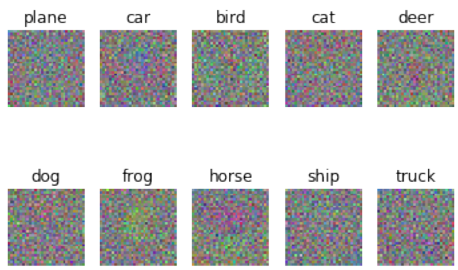

Figure 2:

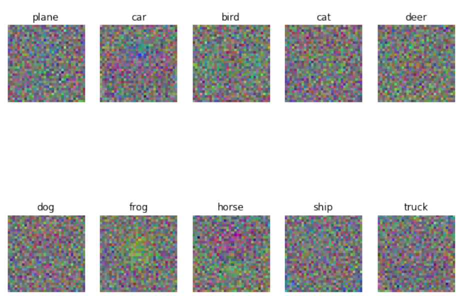

Figure 3:

SVM tends to have higher learning rate and lower regularization strength comparing with Softmax.

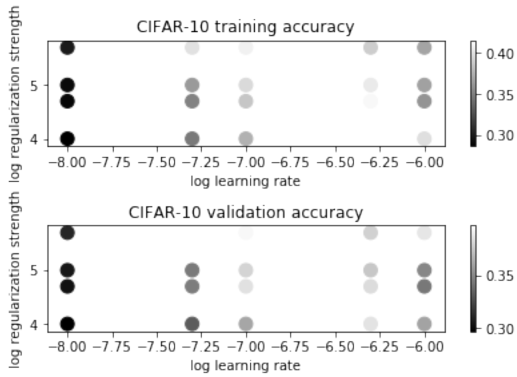
CIFAR-10 training accuracy
log regularization strength
log learning rate
CIFAR-10 validation accuracy
log regularization strength
log learning rate

Figure 4: