FINAL PRACTICE PROBLEMS # 2

Monday, Jun 11st, 2018

Problem 1 (Kernel k-means)

First given a clustering S_i, we will put

$$\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} \phi(x)$$

to minimize $\sum_{x \in S_i} ||\phi(x) - \mu_i||_2^2$
Following this, the optimal clustering is given by assigning x_i to the cluster $arg \min_k f(i,k)$, where

$$\begin{aligned}
f(i,k) &= ||\phi(x_i) - \mu_k||^2 \\
&= \phi(x_i)^T \phi(x_i) - 2\phi(x_i)^T \mu_k + \mu_k^T \mu_k \\
&= \phi(x_i)^T \phi(x_i) - \frac{2}{|S_k|} \sum_{x_j \in S_k} \phi(x_i)^T \phi(x_j) + \frac{1}{|S_k|^2} \sum_{x_j, x_l \in S_k \times S_k} \phi(x_j)^T \phi(x_l) \\
&= K(x_i, x_i) - \frac{2}{|S_k|} \sum_{x_j \in S_k} K(x_i, x_j) + \frac{1}{|S_k|^2} \sum_{x_j, x_l \in S_k \times S_k} K(x_j, x_l)
\end{aligned}$$

Therefore

$$class(i) = rg \min_k \frac{1}{|S_k|^2} \sum_{x_j, x_l \in S_k \times S_k} K(x_j, x_l) - \frac{2}{|S_k|} \sum_{x_j \in S_k} K(x_i, x_j)$$

Problem 2 (Boosting)

On this dataset, there are four nontrivial things that a stump could do:

s_1 classifies the left two points as positive;
s_2 classifies the right two points as positive;
s_3 classifies the top two points as positive;
s_4 classifies the bottom two points as positive.
So the function you end up learning could be anything of the form

$$\hat{y}(x) = \sum_{i=1}^{n} f_i(x)$$

where each f_i is one of the s_j.

Now, note that each copy of s_1 in that sum cancels out a copy of s_2, because they're opposite, and similarly for s_3 and s_4. So $\hat{y}$ is really an integer combination $\hat{y}(x) = as_1(x) + bs_3(x)$

But the first half of that expression doesn't change when you move from top to bottom, and the second half always changes by the same amount (b). So we know that the output of $\hat{y}$ must either always increase as the datapoint moves from top to bottom (if $b < 0$), or always decrease (if $b > 0$).

If it always increases when moving from top to bottom, then it can't get both the top-left and bottom-left points correct (because the top one is greater than 0 and the bottom one is less than 0).

If it always decreases, then similarly it can't get both the top-right and bottom-right points correct.

Therefore, no possible sum of boosted stumps can classify the dataset perfectly,

Problem 3 (SVM)

Recall the soft-margin SVM in the primal:

$$\arg\min_{\mathbf{w},b,\{\xi_n\}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n$$

$$\begin{aligned} y_n(\mathbf{w}^T\mathbf{x}_n + b) &\geq 1 - \xi_n \quad n \in \{1,\ldots,N\} \\ \xi_n &\geq 0 \quad n \in \{1,\ldots,N\} \end{aligned}$$

(a) α_n represents the dual variable associated with constraint n. The support vectors are the datapoints n such that the optimal values of α_n^*, $\alpha_n^* > 0$. Thus, we have $\alpha_1^* > 0$, $\alpha_2^* > 0$ and $\alpha_n^* = 0$ for $n = \{3,\ldots,N\}$.

The KKT conditions (complementary slackness) require that $\alpha_n^*\left(1 - \xi_n - y_n(\mathbf{w}^{*T}\mathbf{x}_n + b^*)\right) = 0$ and $(C - \alpha_n^*)\xi_n = 0$.

For $n \in \{3,\ldots,N\}$, we have $\alpha_n^* = 0$ so that $\xi_n = 0$. For support vectors $n \in \{1,2\}$, we have $\alpha_n^* > 0$ so that $\xi_n = 1 - y_n(\mathbf{w}^{*T}\mathbf{x}_n + b^*)$.

Answers based on intuition are also acceptable, *i.e.*, that slack is zero for non support vectors.

(b) i. Decreases

ii. Decreases

Problem 4 (Kernel)

First we expand the dot product inside, and square the entire sum. We will get a sum of the squares of the components and a sum of the cross products.

$$\begin{aligned} (\mathbf{x}^T\mathbf{y} + c)^2 &= (c + \sum_{i=1}^{n} x_i y_i)^2 \\ &= c^2 + \sum_{i=1}^{n} x_i^2 y_i^2 + \sum_{i=2}^{n}\sum_{j=1}^{i-1} 2x_i y_i x_j y_j + \sum_{i=1}^{n} 2x_i y_i c \end{aligned}$$

Pulling this sum into a dot product of x components and y components, we have

$$\Phi(x) = [c, x_1^2, \cdots, x_n^2, \sqrt{2}x_1x_2, \cdots, \sqrt{2}x_1x_n, \sqrt{2}x_2x_3, \cdots, \sqrt{2}x_{n-1}x_n, \sqrt{2c}x_1, \cdots, \sqrt{2c}x_n]$$

In this feature mapping, we have c, the squared components of the vector $\mathbf{x}$, $\sqrt{2}$ multiplied by all of the cross terms, and $\sqrt{2c}$ multiplied by all of the components.

Problem 5 (Logistic Regression and Perceptron (28 pts))

(a) From the update rule $\nabla_w J(\boldsymbol{w}) = \sum_{i=1}^{n}(h_{\boldsymbol{w}}(\mathbf{x}_i) - y_i)\boldsymbol{x}_i$.
From the expression for the gradient you can see that

$$\frac{\partial J(\boldsymbol{w})}{\partial w_j} = \sum_{i=1}^{n}(h_{\boldsymbol{w}}(\boldsymbol{x}_i) - y_i)x_{i,j}$$

$$\begin{aligned}\frac{\partial^2 J(\boldsymbol{w})}{\partial w_k \partial w_j} &= \frac{\partial}{\partial w_k}\Big(\sum_{i=1}^{n}(h_{\boldsymbol{w}}(\boldsymbol{x}_i) - y_i)x_{i,j}\Big) \\ &= \sum_{i=1}^{n}\frac{\partial}{\partial w_k}h_{\boldsymbol{w}}(\boldsymbol{x}_i)x_{i,j} \\ &= \sum_{i=1}^{n}h_{\boldsymbol{w}}(\boldsymbol{x}_i)(1 - h_{\boldsymbol{w}}(\boldsymbol{x}_i))x_{i,j}x_{i,k}\end{aligned}$$

Therefore we have

$$\nabla_w^2 J(\boldsymbol{w}) = \sum_{i=1}^{n}h_{\boldsymbol{w}}(\boldsymbol{x}_i)(1 - h_{\boldsymbol{w}}(\boldsymbol{x}_i))\boldsymbol{x}_i\boldsymbol{x}_i^T = \boldsymbol{X^T D X}$$

$$\boldsymbol{u^T X^T D X u} = ||D^{\frac{1}{2}}Xu||_2^2 > 0 \quad \forall \boldsymbol{u} \neq 0$$

(b)

$$\boldsymbol{w}^{t+1} \leftarrow \boldsymbol{w}^t - \eta_t(h_{\boldsymbol{w}^t}(\boldsymbol{x}_{i(t)}) - y_{i(t)})\boldsymbol{x}_{i(t)}$$

Here $i(t) \sim \text{Uniform}[1, \ldots, n]$

(c)

$$\boldsymbol{w}^{t+1} \leftarrow \boldsymbol{w}^t - \eta_t(\hat{y}_i^t - y_i)\boldsymbol{x}_i$$

When $\hat{y}_i^t = y_i$, no update takes place, whereas when $\hat{y}_i^t - y_i = 1$ or -1 the corresponding update takes place. If we use new labels $z_i \in \{-1, 1\}$ instead of $y_i \in \{0, 1\}$ then the update rule becomes

$$\boldsymbol{w}^{t+1} \leftarrow \boldsymbol{w}^t + 2\eta_t z_i \boldsymbol{x}_i$$

For $\eta = 0.5$ we have the Perceptron algorithm as discussed in class.

Problem 6 (EM algorithm and Gaussian Mixture Model)

The new estimates will be

$$
\begin{aligned}
w_1 &= (0.2 + 0.2 + 0.8 + 0.9 + 0.9)/5 = 0.6 \\
w_2 &= (0.8 + 0.8 + 0.2 + 0.1 + 0.1)/5 = 0.4 \\
\mu_1 &= (0.2 \times 5 + 0.2 \times 15 + 0.8 \times 25 + 0.9 \times 30 + 0.9 \times 40)/(0.2 + 0.2 + 0.8 + 0.9 + 0.9) = 29 \\
\mu_2 &= (0.8 \times 5 + 0.8 \times 15 + 0.2 \times 25 + 0.1 \times 30 + 0.1 \times 40)/(0.8 + 0.8 + 0.2 + 0.1 + 0.1) = 14
\end{aligned}
$$