Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

QUESTION

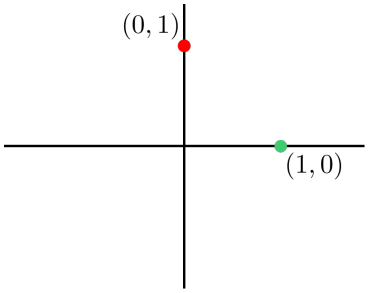
1

Assignment Number: 1
Student Name: Varun Khare
Roll Number: 150793
Date: September 10, 2017

1 Voronoi for U

The distance between the points is defined as $d(\mathbf{z}^1, \mathbf{z}^2) = \langle \mathbf{z}^1 - \mathbf{z}^2, U(\mathbf{z}^1 - \mathbf{z}^2) \rangle$ where

$$\mathbf{U} = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$$



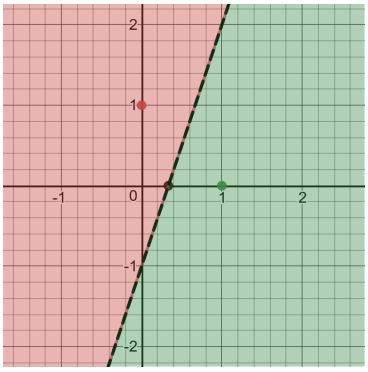


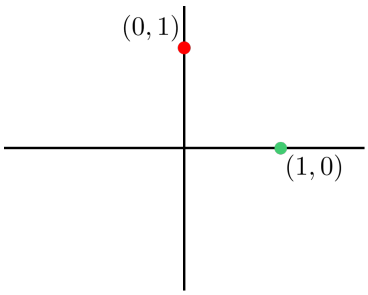
Figure 1: Learning with Prototypes: the figure on the left shows the two prototypes. The figure on the right shows what the decision boundary if the distance measure used is $d(\mathbf{z}^1, \mathbf{z}^2)$ as defined above, for any two points $\mathbf{z}^1, \mathbf{z}^2 \in \mathbb{R}^2$.

The decision boundary in this case is the line $\mathbf{y = 3x - 1}$.

2 Voronoi for V

The distance between the points is defined as $d(\mathbf{z}^1, \mathbf{z}^2) = \langle \mathbf{z}^1 - \mathbf{z}^2, U(\mathbf{z}^1 - \mathbf{z}^2) \rangle$ where

$$= \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$



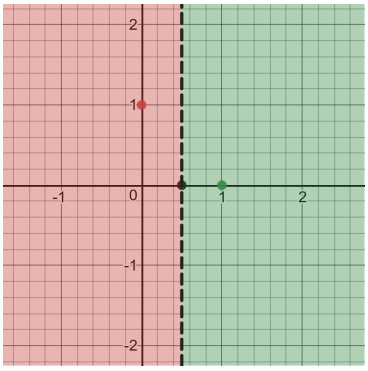


Figure 2: Learning with Prototypes: the figure on the left shows the two prototypes. The figure on the right shows what the decision boundary if the distance measure used is $d(\mathbf{z}^1, \mathbf{z}^2)$ as defined above, for any two points $\mathbf{z}^1, \mathbf{z}^2 \in \mathbb{R}^2$.

Definition 1.1. The decision boundary in this case is the line $\mathbf{2x = 1}$.

Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

Assignment Number: 1
Student Name: Varun Khare
Roll Number: 150793
Date: September 10, 2017

QUESTION

2

3 Likelihood Distribution

The solution to the least-squares regression problem on the dataset $\left\{\mathbf{x}^i, y^i\right\}_{i=1,2,3\ldots n}$, where $\mathbf{x}^i \in \mathbb{R}^d$ and $y \in \mathbb{R}$ is given by $\hat{\mathbf{w}}_{\text{cls}}$. The likelihood distribution, with variance σ^2, ($\mathbb{P}[\mathbf{y}|X, \mathbf{w}]$) is given by

Proposition 1.1.

$$\mathbb{P}[y|x^i, \mathbf{w}] \sim \mathcal{N}(\left\langle \mathbf{w}, \mathbf{x}^i \right\rangle, \sigma^2)\,.$$

$$\mathbb{P}[y^i|\mathbf{x}^i, \mathbf{w}] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(y^i - \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)^2}{2\sigma^2} \right)$$

$$\mathbb{P}[\mathbf{y}|X, \mathbf{w}] = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(y^i - \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)^2}{2\sigma^2} \right)$$

4 Prior Distribution

The prior distribution is the probability $\mathbb{P}[\mathbf{w}]$. For this case it is equal to

Proposition 1.2.

$$\mathbb{P}[\mathbf{w}] = \begin{cases} k \text{ when } \|\mathbf{w}\|_2 \leq r \\ 0 \text{ otherwise} \end{cases}$$

Here k is the area covered by the points lying in $\|\mathbf{w}\|_2 \leq r$ i.e. $k = \int_{\|\mathbf{w} \leq r\|} d\mathbf{w}$

Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

Assignment Number: 1
Student Name: Varun Khare
Roll Number: 150793
Date: September 10, 2017

QUESTION

3

5 Likelihood Distribution

The solution to the feature regularized least-squares regression problem on the data set $\left\{\mathbf{x}^i, y^i\right\}_{i=1..n}$, where $\mathbf{x}^i \in \mathbb{R}^d$ and $y \in \mathbb{R}$ is given by $\hat{\mathbf{w}}_{\text{fr}}$. The likelihood distribution, with variance σ^2, ($\mathbb{P}[\mathbf{y}|X, \mathbf{w}]$) is again a normal distribution characteristic of the least-squares minimzer.

Proposition 1.3.

$$\mathbb{P}[y|x^i, \mathbf{w}] \sim \mathcal{N}(\left\langle \mathbf{w}, \mathbf{x}^i \right\rangle, \sigma^2)\,.$$

$$\mathbb{P}[y^i|\mathbf{x}^i, \mathbf{w}] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(y^i - \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)^2}{2\sigma^2} \right)$$

$$\mathbb{P}[\mathbf{y}|X, \mathbf{w}] = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(y^i - \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)^2}{2\sigma^2} \right)$$

6 Prior Distribution

The prior distribution is the probability $\mathbb{P}[\mathbf{w}]$. Instead of the usual L_2 regularizer we have a custom feature regularizer. Just like the standard regression prior we can get the prior for this model by tasking the negative exponentiation of the regularizer. (This can be easily proved)

Proposition 1.4.

$$\mathbb{P}[\mathbf{w}] = \frac{\sigma^d}{\sqrt{(2\pi)^{\frac{n}{2}} (\prod_{i=1}^{d} \alpha_i)}} \exp\left(- \frac{\sum_{i=1}^{n} \alpha_i \mathbf{w}_i^2}{2\sigma^2} \right)$$

σ is some constant there. It can be evaluated by taking the integral of $\mathbb{P}[\mathbf{w}]$ over entire space and equating to 1.

7 MAP Estimate Closed Form Solution

By equating the derivative at minima equal to 0 we get, $\frac{\partial f}{\partial w} = 0$

$$\Rightarrow -\mathbf{X}^T \cdot \mathbf{y} + \mathbf{X}^T\mathbf{X}\hat{\mathbf{w}}_{\text{fr}} + \mathbf{K}\hat{\mathbf{w}}_{\text{fr}} = 0$$

Here K is the diagonal matrix with $K : \{K_{ii} = \alpha_i\}$. Hence we get $\hat{\mathbf{w}}_{\text{fr}} = (\mathbf{X}^T\mathbf{X} + \mathbf{K})^{-1}\mathbf{X}^T\mathbf{y}$. The symbols are in accordance to problem defined.

To prove this equation we will first prove that optimal solution of $P2$ is a solution of $P1$.

Lemma 1.5. If $\{\hat{\mathbf{W}}^0, \hat{\epsilon}^0\}$ is the solution of $P1$ then its is also the solution of the $P2$

7.1 Proof:

Assume $\hat{\mathbf{W}}^0$ is not the solution of the $P2$. Then $\exists\ \widehat{\mathbf{W}}$ st

$$P2(\widehat{\mathbf{W}}^0) \geq P2(\widehat{\mathbf{W}}) \tag{1}$$

Now,

$$l_{cs}(y^i, \eta^i) = [1 + \underset{k \neq y}{max}\ \eta_k^i - \eta_y^i]_+ \tag{2}$$
$$= [\underset{k \neq y}{max}\ 1 + \eta_k^i - \eta_y^i]_+ \tag{3}$$
$$= \underset{k \neq y}{max}[1 + \ \eta_k^i - \eta_y^i]_+ \tag{4}$$
$$\tag{5}$$

Let Θ be equal to $l_{cs}(y^i, \eta^i)$ then,

$$\Theta \geq 1 + \ \eta_k^i - \eta_y^i\ \forall k \neq y^i \tag{6}$$
$$\Rightarrow \Theta^i \geq 1 + \left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle\ - \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle, \forall i, \forall k \neq y^i \tag{7}$$

We already know that $\Theta^i \geq 0$, replacing Θ^i with l_{cs} we get $P1$. So $P2$ gets converted to $P1$ at $\widehat{\mathbf{W}}$ and by assumption $\widehat{\widehat{\mathbf{W}}}$ gives a lower value than $\widehat{\mathbf{W}}^0$ which means $\widehat{\mathbf{W}}^0$ is not the optimal solution. Hence, by contradiction the lemma is proved.
We can also simultaneously derive $P2$ directly from $P1$ by following the same equations. So, the solutions of $P2$ are subset of solutions of $P1$. From this and lemma we prove the equality of the sets of the solutions of $P1$ and $P2$. Thus, the two problems are equivalent.

Theorem 1.6. $f(\mathbf{w}') \geq f(\mathbf{w}) + \langle \mathbf{g}, \mathbf{w}' - \mathbf{w} \rangle$

7.2 Proof:

the problem can be rewritten as

$$\sum_{i=1}^{n} \left([1 - \mathbf{y}^i \langle \mathbf{w}', \mathbf{x}^i \rangle]_+ - [1 - \mathbf{y}^i \langle \mathbf{w}, \mathbf{x}^i \rangle] \right) \geq \left\langle \sum_{i=1}^{n} \mathbf{h}^i, \mathbf{w}' - \mathbf{w} \right\rangle \tag{8}$$

$$\geq \sum_{i=1}^{n} \langle \mathbf{h}^i, \mathbf{w}' - \mathbf{w} \rangle \tag{9}$$

Now, comparing the i^{th} terms on both sides we have:

$$\textbf{L.H.S term} = \begin{cases} 0 & when\ 1 \leq \mathbf{y}^i \langle \mathbf{w}', \mathbf{x}^i \rangle\ \&\ 1 \leq \mathbf{y}^i \langle \mathbf{w}, \mathbf{x}^i \rangle \\ \mathbf{y}^i \mathbf{w}^T \mathbf{x}^i - 1 & when\ 1 \leq \mathbf{y}^i \langle \mathbf{w}', \mathbf{x}^i \rangle\ \&\ 1 \geq \mathbf{y}^i \langle \mathbf{w}, \mathbf{x}^i \rangle \\ 1 - \mathbf{y}^i \mathbf{w}'^T \mathbf{x}^i & when\ 1 \geq \mathbf{y}^i \langle \mathbf{w}', \mathbf{x}^i \rangle\ \&\ 1 \leq \mathbf{y}^i \langle \mathbf{w}, \mathbf{x}^i \rangle \\ \mathbf{y}^i (\mathbf{w}^T - \mathbf{w}'^T) \mathbf{x}^i & when\ 1 \geq \mathbf{y}^i \langle \mathbf{w}', \mathbf{x}^i \rangle\ \&\ 1 \geq \mathbf{y}^i \langle \mathbf{w}, \mathbf{x}^i \rangle \end{cases}$$

Utilizing the definitions of $\mathbf{h}^i$ to evaluate the i^{th} RHS term we get

$$\textbf{R.H.S term} = \begin{cases} 0 & when\ 1 \leq \mathbf{y}^i \langle \mathbf{w}', \mathbf{x}^i \rangle\ \&\ 1 \leq \mathbf{y}^i \langle \mathbf{w}, \mathbf{x}^i \rangle \\ -\mathbf{y}^i (\mathbf{w}'^T - \mathbf{w}^T) \mathbf{x}^i - 1 & when\ 1 \leq \mathbf{y}^i \langle \mathbf{w}', \mathbf{x}^i \rangle\ \&\ 1 \geq \mathbf{y}^i \langle \mathbf{w}, \mathbf{x}^i \rangle \\ 0 & when\ 1 \geq \mathbf{y}^i \langle \mathbf{w}', \mathbf{x}^i \rangle\ \&\ 1 \leq \mathbf{y}^i \langle \mathbf{w}, \mathbf{x}^i \rangle \\ -\mathbf{y}^i (\mathbf{w}'^T - \mathbf{w}^T) \mathbf{x}^i & when\ 1 \geq \mathbf{y}^i \langle \mathbf{w}', \mathbf{x}^i \rangle\ \&\ 1 \geq \mathbf{y}^i \langle \mathbf{w}, \mathbf{x}^i \rangle \end{cases}$$

Direct Comparison of LHS and RHS terms brings $LHS \geq RHS$ for a given $\mathbf{y}^i, \mathbf{x}^i, \mathbf{w}, \mathbf{w}'$. Thus i^{th} term of left summation always greater than the i^{th} term of right summation. Thus adding all these terms we get $f(\mathbf{w}') \geq f(\mathbf{w}) + \langle \mathbf{g}, \mathbf{w}' - \mathbf{w} \rangle$. Hence proved.

8 Testing Observations

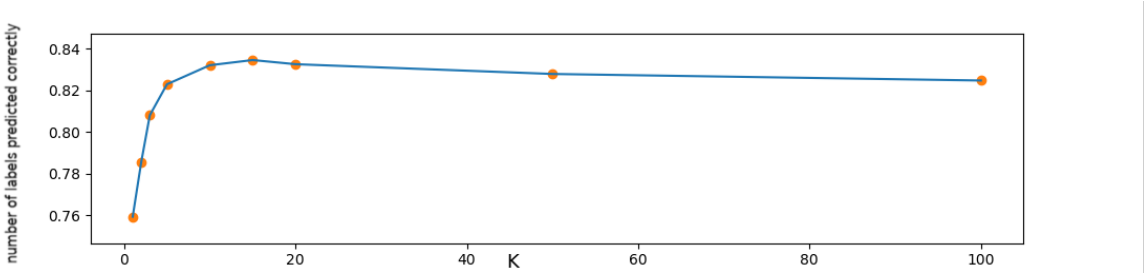


Figure 3: As you can see from the graph the accuracy increases with the value of k upto k 15. After this the accuracy is fairly constant and then starts dropping.

T he reason for this is in the beginning (when k=1) the model is over fitting. Even a slight change in the distances of the nearest neighbor can cause the model to give different predictions. So the model accuracy drops to less robustness to noise in the data.

A fter a peek stage, the accuracy again starts dropping. this is the case of under fitting. In these cases the model starts taking several points for consideration. Now these points can be very far apart and still be one of the k nearest neighbors. Since they are very far they have very different feature values and and hence may represent different type of data but large k ignores these facts. its like when k is sufficiently (of the order of size of the data set) the model gives same prediction for every point based on the majority class. In this extreme case it has lost the entire information about data distribution.

9 K validation results

With the experiments during 20000 fold cross validation I got the (K,accuracy) pairs as {(10,0.830366), (12,0.83208), (16,0.832), (18,0.8319)}.

Clearly the optimal choice of k for me is **16**.

10 LMNN

The accuracy achieved on the test set is 83.41% for the learnt Mahalanobis matrix and k=16.