CS446 HW1

Junsheng Huang

January 25, 2024

1

1.1

For each test image, we need to calculate the distance for every dataset node in worst case. We have M test images and N training images and each distance calculate need d dimensions, so the answer is O(M*N*d)

1.2

k = 10 should have a "smoother" decision boundary. Because for bigger k, we would have a simpler model and more training nodes to "vote" for single test node. For example, when the k is too big to contain all the training set, the decision boundary is the whole set since we only have one decision. When k = 1, the model is over-fit and decision boundary is made up of many single nodes.

1.3

Consider $\mathbf{x} = \begin{bmatrix} x1 \\ x2 \end{bmatrix}$, $\mathbf{w} = [1, 1]$, $b = 0$, so $\mathbf{w} \cdot \mathbf{x} + b = x1 + x2$ and this matches ground truth g.

1.4

The relationship between the largest singular value of A (named it as x) and the largest eigenvalue of $A^T A$ (named it as λ) is $x = \sqrt{\lambda}$.
This is because: (remember V is orthogonal matrix)

$$\begin{aligned} A^T \cdot A &= (U\Sigma V^T)^T (U\Sigma V^T) && (1) \\ &= V\Sigma^T U^T \cdot U\Sigma V^T && (2) \\ &= V\Sigma^T \Sigma V^T && (3) \\ &= V\Sigma^T \Sigma V^{-1} && (4) \end{aligned}$$

That means Σ^2 is the eigenvalue matrix of $A^T A$, so the relationship is $x = \sqrt{\lambda}$.

1.5

The Naive Bayes assumption is violated at text world. For example, the sequence "of course" is more common than it would be if the words "of" and "course" were conditionally independent. That is, $Pr(X = ofcourse|Y) > Pr(x1 = of|Y)Pr(x2 = course|Y)$.

2

2.1

We can write $X \in R^{n \times d}$ to $A \in R^{d \times d}$,$w^{'} \in R^d$, $B \in R^d$, all new rows are filled with 0, so the rank of A is n, which is not full-rank, thus $w^{'}$ has at least one solution, and if there exist $w^{'}$ that satisfy $Aw^{'} = B$, there must exist w that satisfy $Xw = y$, and the empirical risk with squared loss is zero. (Or you can consider that the dimension of X 's column space is n, and $\mathbf{y} \in R^n$, so $\mathbf{y}$ is in the column space of X, thus we can say that there always exist a $\mathbf{w}$ that $X\mathbf{w} = \mathbf{y}$.)

2.2

The rank of x is n, so we have n eigenvalue and singular value. Since Σ should have the singular value at its diagonal, the rank of Σ is n.

2.3

$$X \cdot X^T = (U\Sigma V^T)(U\Sigma V^T)^T \tag{5}$$
$$= U\Sigma V^T \cdot V\Sigma^T U^T \tag{6}$$
$$= U\Sigma\Sigma^T U^T \tag{7}$$

since U, Σ^2 and U^T are all full-rank matrices, XX^T is all full-rank, thus is invertible.

3

3.1

The smallest number of support vectors is 2. This happens when exactly 2 support vectors on the hard margin of the SVM.

3.2

since the optimal solution to the dual only has 3 values that is not zero, there are at least 3 support vectors. So the smallest number of support vectors is 3, the largest number of support vectors is the total number of α (or 10000 if using first part's data number).

3.3

3.3.1

$$k(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^T\boldsymbol{z} + 1)^2 \tag{8}$$
$$= (x_1z_1 + x_2z_2 + 1)^2 \tag{9}$$
$$= x_1^2z_1^2 + x_2^2z_2^2 + 2x_1z_1x_2z_2 + 2x_1z_1 + 2x_2z_2 + 1 \tag{10}$$
$$= \phi(\boldsymbol{x})^T\phi(\boldsymbol{z}) \tag{11}$$

so for $\boldsymbol{x} = (x_1, x_2)$, $\phi(\boldsymbol{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1)$

3.3.2

since we only consider the sign of x_1 and x_2 (same for 1 and different for -1): we can only consider the $\sqrt{2}x_1x_2$ in $\phi(\boldsymbol{x})$, which means $w = (0, 0, 1, 0, 0, 0)$ can work.

4

4.1

$$P(y = +1|x) = \frac{P(x, y = +1)}{P(x)} \tag{12}$$
$$= \frac{P(x|y = +1) \cdot P(y = +1)}{P(x|y = +1) \cdot P(y = +1) + P(x|y = -1) \cdot P(y = -1)} \tag{13}$$
$$= \frac{P(x|y = +1) \cdot p}{P(x|y = +1) \cdot p + P(x|y = +1) \cdot (1 - p)} \tag{14}$$
$$= \frac{1}{1 + \frac{P(x|y=-1)\cdot(1-p)}{P(x|y=+1)\cdot p}} \tag{15}$$
$$= \frac{1}{1 + exp(log\frac{A}{B})} \tag{16}$$
$$\tag{17}$$

where $A = P(x|y = -1) \cdot (1 - p)$, $B = P(x|y = +1) \cdot p$

4.2

assumption: each x_j is conditional independent, $x = \begin{bmatrix} x_1 \\ ... \\ x_n \end{bmatrix}$

Thus, we have:

$$log(\frac{A}{B}) = log(\frac{P(x|y=-1)\cdot(1-p)}{P(x|y=+1)\cdot p}) = log(\frac{1-p}{p}) + log(\frac{P(x|y=-1)}{P(x|y=+1)}) \tag{18}$$

$$log(\frac{P(x|y=-1)}{P(x|y=+1)}) = log(exp(-\frac{1}{2}\cdot\Sigma(x_j-\mu_{-,j})^2 + \frac{1}{2}\cdot\Sigma(x_j-\mu_{+,j})^2)) \tag{19}$$

$$= -\frac{1}{2}\Sigma(\mu_{-,j})2 + \frac{1}{2}\Sigma(\mu_{+,j})2 + \Sigma x_j(\mu_{-,j}-\mu_{+,j}) \tag{20}$$

so $log(\frac{A}{B})$ can be written in the form $\boldsymbol{w}^T\boldsymbol{x}+b$, $\boldsymbol{w} = \boldsymbol{\mu_-} - \boldsymbol{\mu+}$, $b = log(\frac{1-p}{p}) + \frac{1}{2}\boldsymbol{\mu_+}\boldsymbol{\mu_+}^T - \frac{1}{2}\boldsymbol{\mu_-}\boldsymbol{\mu_-}^T$.

4.3

$$P(y|\boldsymbol{x}) = \frac{1}{1+exp(y\cdot(\boldsymbol{w}^T\boldsymbol{x}+b))} \tag{21}$$

5

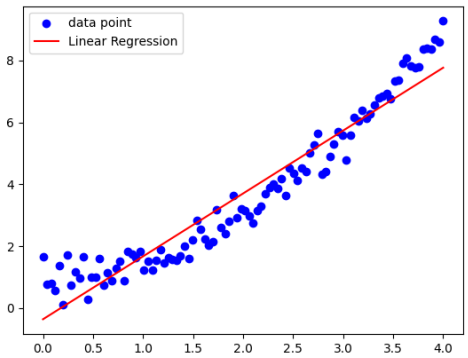

Figure 1: Linear Regression