Bingjun Guo (bingjun3)

0 Instructions

Homework is due Thursday, February 6, 2024 at 23:59pm Central Time. Please refer to `https://courses.grainger.illinois.edu/cs446/sp2024/homework/hw/index.html` for course policy on homeworks and submission instructions.

1 Short answer: 10pts

1. $O(MNd)$
2. $k = 10$
3. $\left(\begin{bmatrix}1\\1\end{bmatrix}, 1\right)$
4. The largest eigenvalue of $A^\top A$ is the square of the largest singular value of A.
5. In sentiment recognization tasks on natural language, for example, movie comments, for the probability of the phrase "not good" appearing in a positive comment:

$$P(\text{“}not\text{”}, \text{“}good\text{”}|positive) \neq P(\text{“}not\text{”}|positive) \cdot P(\text{“}good\text{”}|positive)$$

since both $P(\text{“}not\text{”}|positive)$ and $P(\text{“}good\text{”}|negative)$ are adequately high but "not good" should appear really rare in positive comments.

2 Linear Regression: 10pts

1. X can be considered as a linear transform from $\mathbb{R}^n$ to $\mathbb{R}^d$. Thus, it complies to the Rank-Nullity Theorem:

$$\text{rank}(X) + \text{nullity}(X) = \dim(\text{domain}(X))$$

in which the dimention of the input space for X is n and $\text{rank}(X) = n$.
Therefore, $\text{nullity}(X) = 0$, which indicates that X is invertible. Thus, there exists $\boldsymbol{w} = X^{-1}\boldsymbol{y}$ such that $X\boldsymbol{w} = \boldsymbol{y}$.

2. Since the number of non-zero singular values of A equals to $\text{rank}(A)$, Σ is a diagonal matrix consists of positive singular values of A, and X is real, $\text{rank}(\Sigma) = \text{rank}(A) = n$.

3. Firstly we will prove that $X^\top$ and $XX^\top$ share the same nullity,
   i.e., $X^\top M = 0 \iff XX^\top M = 0$ for $M \in \mathbb{R}^n$.
   $X \cdot 0 = 0$, thus $X^\top M = 0 \to X(X^\top M) = 0 \to XX^\top M = 0$.
   Supoose $XX^\top M = 0$, then we have $M^\top XX^\top M = 0$, and thus $(X^\top M)^\top X^\top M = 0$, with $X^\top M \in \mathbb{R}^d$. $(X^\top M)^\top X^\top M$ equals to sum of square of all the entries in $X^\top M$, which can only be greater or equal to 0 since its a real vector. Thus, all entries in $X^\top M$ are 0, *i.e.*, $X^\top M = \mathbf{0}$. Therefore, $XX^\top M = 0 \to X^\top M = 0$.
   Secondly, since $\text{nullity}(X^\top) = \text{nullity}(XX^\top)$, according to the Rank-Nullity Theorem introduced in the first question, since the dimention of input space (RHS of the equation) is both n for linear transforms $X^\top : \mathbb{R}^n \to \mathbb{R}^d$ and $XX^\top : \mathbb{R}^n \to \mathbb{R}^n$, $\text{rank}(XX^\top) = \text{rank}(X^\top) = \text{rank}(X) = n$. Therefore, $XX^\top$ is a full-rank square matrix, and thus $XX^\top$ is invertible.

3 SVM: 10 pts

1. 2, which happens in the case that the closest two vectorsa in different class are selected and there's only such 2 points in $\mathcal{D}$ belonging to different classes that have such distance between each other.

2. With optimal $\boldsymbol{w}^* = \sum_{i\in[n]} \alpha_i^* y_i \boldsymbol{x}_i$, $(\boldsymbol{x}_i, y_i)$ is a support vector if and only if $y_i \left(\sum_{j\in[n]} \alpha_j^* y_j \boldsymbol{x}_j^\top\right) \boldsymbol{x}_i = 1$, that is, $\sum_{j\in[n]} \alpha_j^* (y_j \boldsymbol{x}_j)^\top = (\boldsymbol{X}\boldsymbol{\alpha}^*)^\top = \frac{\boldsymbol{x}_i^{-1}}{y_i}$, in which both $\boldsymbol{X}$ and $\frac{\boldsymbol{x}_i^{-1}}{y_i}$ are fixed, while $\boldsymbol{\alpha}^*$ serve as a linear combination might be with multiple optimal solutions. It's possible that this mapping stands while $\alpha_i = 0$, in which case $\boldsymbol{x}_i$ is indeed a support vector while hasn't been observed. However, if the vector $\boldsymbol{x}_i$ is observed with a non-zero α_i, it's sure to be a support vector. Therefore, the smallest possible number of support vectors in $\mathcal{D}$ is 3 and the largest possible number of support vectors in $\mathcal{D}$ would be n.

3. (a)
   $$\phi(\boldsymbol{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1)$$
   (b)
   $$\phi(-1,-1) = (1, 1, \sqrt{2}, -\sqrt{2}, -\sqrt{2}, 1)$$
   $$\phi(1,1) = (1, 1, \sqrt{2}, \sqrt{2}, \sqrt{2}, 1)$$
   $$\phi(1,-1) = (1, 1, -\sqrt{2}, \sqrt{2}, -\sqrt{2}, 1)$$
   $$\phi(-1,1) = (1, 1, -\sqrt{2}, -\sqrt{2}, \sqrt{2}, 1)$$
   Therefore, $\boldsymbol{w}$ can be $(0, 0, 1, 0, 0, 0)$.

4 Gaussian Naive Bayes: 15pts

1.

$$\frac{1}{1+\exp(\log\frac{A}{B})} = \frac{B}{B+A}$$

$$P(y=+1|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|y=+1)\cdot p}{P(\boldsymbol{x})} = \frac{P(\boldsymbol{x}|y=+1)\cdot p}{P(\boldsymbol{x}|y=+1)\cdot p + P(\boldsymbol{x}|y=-1)\cdot(1-p)}$$

Therefore, for $B = P(\boldsymbol{x}|y=+1)\cdot p$ and $A = P(\boldsymbol{x}|y=-1)\cdot(1-p)$,

$$P(y=+1|\boldsymbol{x}) = \frac{1}{1+\exp(\log\frac{A}{B})}$$

2.

$$P(y=+1|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|y=+1)\cdot p}{P(\boldsymbol{x})}$$

$$= \frac{\prod_{i=1}^{d} P(x_i|y=+1)\cdot p}{\prod_{i=1}^{d} P(x_i|y=+1)\cdot p + \prod_{i=1}^{d} P(x_i|y=-1)\cdot(1-p)}$$

$$= \frac{1}{1+\exp\left(\log\frac{\prod_{i=1}^{d} P(x_i|y=-1)\cdot(1-p)}{\prod_{i=1}^{d} P(x_i|y=+1)\cdot p}\right)}$$

Since $P(x_j|y=+1) = \frac{1}{\sqrt{2\pi}}\exp(\frac{-(x_j-\mu_{+,j})^2}{2})$, we have:

$$\frac{P(x_j|y=-1)}{P(x_j|y=+1)} = \exp\left(\frac{-(x_j-\mu_{-,j})^2+(x_j-\mu_{+,j})^2}{2}\right)$$

and thus:

$$\log\frac{A}{B} = \log\frac{\prod_{i=1}^{d} P(x_i|y=-1)\cdot(1-p)}{\prod_{i=1}^{d} P(x_i|y=+1)\cdot p}$$

$$= \log \frac{1-p}{p} + \frac{1}{2} \sum_{i=1}^{d} -(x_i - \mu_{-,i})^2 + (x_i - \mu_{+,i})^2$$

$$= \log \frac{1-p}{p} + \frac{1}{2} \sum_{i=1}^{d} (2x_i - \mu_{-,i} - \mu_{+,i})(\mu_{-,i} - \mu_{+i})$$

$$= \log \frac{1-p}{p} + \sum_{i=1}^{d} x_i \cdot (\mu_{-,i} - \mu_{+,i}) - \frac{1}{2} \sum_{i=1}^{d} (\mu_{-,i} + \mu_{+,i})(\mu_{-,i} - \mu_{+,i})$$

$$= (\boldsymbol{\mu}_-^\top - \boldsymbol{\mu}_+^\top)\boldsymbol{x} + \frac{1}{2}(\boldsymbol{\mu}_+^\top \boldsymbol{\mu}_+ - \boldsymbol{\mu}_-^\top \boldsymbol{\mu}_-) + \log \left(\frac{1}{p} - 1 \right)$$

$$= \boldsymbol{w}^\top \boldsymbol{x} + b$$

with $\boldsymbol{w} = \boldsymbol{\mu}_- - \boldsymbol{\mu}_+$ and $b = \frac{1}{2}(\boldsymbol{\mu}_+^\top \boldsymbol{\mu}_+ - \boldsymbol{\mu}_-^\top \boldsymbol{\mu}_-) + \log \left(\frac{1}{p} - 1 \right)$.

3.

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

5 Linear regression: 14pts + 1pt