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. $\forall \hat{x} \in D_{\text{test}}$, where $D_{\text{test}} = \{(x^{(j)})_{j=1}^{N}\}$, we need to compute $||\hat{x} - x^{(i)}||_2 \ \forall i \in [N]$. Since $\hat{x}, x \in R^d$, the l_2 norm takes $O(d)$ time and doing so for all N images requires $O(N \times d)$ time. Additionally, we do this operation for all M test images which in total requires $\boxed{O(M \times N \times d)}$ time.
2. A smaller number of neighbors creates boundaries that are more sensitive to changes in data, hence it can create more jagged boundaries. Also, with a small number of neighbors, there may be a chance that outlier data can dominate as neighbors, to create boundaries that are not smooth. Hence, a larger k value is expected to yield a smoother boundary. So, $\boxed{k = 10}$ is the answer.
3. $\boxed{w = [1, 1], b = 0 \to ([1, 1], 0)}$
   This works because for $x^T = [-1, -1]$
   $$h(x) = sign(-1 - 1 + 0) = sign(-2) = -1$$
   Otherwise for $x^T = [-1, 1], x^T = [1, -1], x^T = [1, 1]$
   $$h([1, 1]^T) = sign(2) = +1$$
   $$h([-1, 1]^T) = h([1, -1]^T) = sign(0) = +1$$
4. SVD of A is $U\Sigma V^T$ for $A \in R^{n \times m}$.
   For $A^T A$, the SVD is $(U\Sigma V^T)^T U\Sigma V^T = V\Sigma^T U^T U\Sigma V^T = V\Sigma^T \Sigma V^T$.
   $A^T A$ can also have an eigendecomposition $W\Lambda W^T$ for $W = V, \Lambda = \Sigma^T \Sigma$.
   Here, Λ is a diagonal matrix whose entries are the singular values squared, σ_i^2, implying,
   $$\boxed{\lambda_{max}(A^T A) = (\sigma_{max}(A))^2}$$
5. Email spam detection is an example where a model can make wrong assumptions that the dataset is independent of one another. Often words or even groups of words (the

input data in this case) are dependent on other words since they try to convey an overall meaning and proper diction is used based on the overall context.

2 Linear Regression: 10pts

1. Given $X \in R^{n\times d}, d > n$. We want to show that $\exists w \in R^d$ such that $L(w)$ is 0, i.e $L(w) = ||Xw - y||_2^2 = 0$. By property of linear system of equations, $Xw = y$ has a solution iff $y \in Col(X)$ or equivalently that y is a linear combination of the columns of X. Since we are given that X is full rank, then $Rank(X) = n = dim(Col(X))$. This also means that the rank of X matches the rank of the augmented matrix, $[X|Y]$, (which comes from doing Gaussian Elimination). Furthermore, we note that X has more columns (d) than rows (n), hence we have an under-determined system of linear equations, where the number of parameters exceeds the number of available equations. Since X is full rank, then this system has an infinite number of solutions. This means that $\exists w$ such that $Xw = y$.

2. $\boxed{Rank(X) = Rank(\Sigma) = n.}$ This is because the number of non-zero singular values in Σ is the same as the number of linearly independent columns in X.

3. $X \in R^{n\times d}, d > n$ and $Rank(X) = n$. This means that $X^T \in R^{d\times n}$, has a full rank of the number of its columns. To show that XX^T is invertible, we need to show that it shares the same rank as X^T or equivalently it shares the same null space as X^T. Hence we can show this for two different cases.

   Case 1: Let $w \in Nul(XX^T)$,

   $$XX^Tw = 0$$
   $$w^T(XX^Tw) = w^TX(X^Tw) = w^T(0)$$
   $$(X^Tw)^T(X^Tw) = 0$$
   $$||X^Tw||_2^2 = 0$$
   $$X^Tw = 0$$

   This implies that $w \in Nul(X^T)$

   Case 2: Let $w \in Nul(X^T)$,

   $$X^Tw = 0$$
   $$X(X^Tw) = X(0)$$
   $$(XX^T)w = 0$$

   This implies that $w \in Nul(XX^T.)$

We can see that XX^T and X^T share the same null space. This is true because $Rank(X^T) = Rank(XX^T) = n$. Given that X^T is full rank, then XX^T is also full rank which means that it is also invertible.

3 SVM: 10pts

1. Since this is a binary classification problem, we will need at least 2 support vectors (one for each class, for a total of 2 classes) to define the margin for the hard-margin SVM. Hence, $\boxed{2}$ is the smallest number of support vectors needed for dataset D.
2. Start by defining,

$$\alpha_i^* : \text{optimal solution to dual } D(\alpha)$$
$$f_i(w^*) : 1 \text{ - y}^{(i)}(w^*)^T x^{(i)}$$

By complementary slackness we know that for $\alpha_i^*, f_i(w)$,

$$\alpha_i^* \times f_i(w^*) = 0, f_i(w^*) \leq 0, \alpha_i^* \geq 0$$

Algebraically, we see that for $(x^{(i)}, y^{(i)})$ to be a support vector,

$$\alpha_i > 0 \implies f_i(w^*) = 0$$

The converse may not be a valid condition to show that a pair is a support vector, hence the single arrow.
Now, we are given optimal $\alpha^* = [10, 2, 3, 0, ..., 0]$. This has three non-zero elements and the rest are all zeroes. By the definition provided in the question, we can see that the smallest number of support vectors in D is $\boxed{3}$. For the upper bound, we know for sure that there will be at least 3 support vectors, but since we know that the dual solution is not unique and that the support vectors may not be unique from the dual solutions, then we can have a maximum of $\boxed{10,000}$ support vectors at most.

3. (a) Define the feature mapping $\phi : R^2 \to R^d$ for $w \in R^2$ as,

$$\boxed{\phi(w) = [1, \sqrt{2}w_1, \sqrt{2}w_2, \sqrt{2}w_1w_2, w_1^2, w_2^2]^T}$$

Where $\boxed{d = 6.}$
Evaluating,

$$\phi(x)^T\phi(z) = 1 + 2x_1z_1 + 2x_2z_2 + 2x_1z_1x_2z_2 + (x_1z_1)^2 + (x_2z_2)^2$$

Rewriting,

$$\phi(x)^T\phi(z) = (x_1z_1 + x_2z_2)^2 + 2(x_1z_1 + x_2z_2) + 1$$

Simplifying,

$$\phi(x)^T\phi(z) = ((x_1z_1 + x_2z_2) + 1)^2$$

$$\phi(x)^T\phi(z) = (\sum_{i=1}^{2}(x_iz_i) + 1)^2$$

Recognizing this as a dot product,

$$\phi(x)^T\phi(z) = (x^Tz + 1)^2$$

(b) Define $w \in R^6$ as such,

$$\boxed{w = [0, 0, 0, 1, 0, 0]^T}$$

Then,

$$sign(w^T\phi(x)) = sign(\sqrt{2}x_1x_2)$$

For $x = [1, 1]$ or $[-1, -1]$,

$$sign(w^T\phi(x)) = sign(\sqrt{2} \times 1^2) = +1$$

This matches $g_{XNOR}(x) = +1$.
Similarly for $x = [1, -1]$ or $[-1, 1]$,

$$sign(w^T\phi(x)) = sign(\sqrt{2} \times -1) = -1$$

This matches $g_{XNOR}(x) = -1$.

4 Gaussian Naive Bayes: 15pts

1. Rewrite the predictor using Bayes Formula,

$$P(y = +1|x) = \frac{P(x|y = +1) \times P(y = +1)}{P(x)}$$

We are given,

$$P(y = +1) = p, p \in (0, 1)$$

Rewriting $P(x)$ using marginal pmf definition,

$$P(x) = \sum_{z \in \{-1,1\}} P(x|y=z) \times P(y=z)$$

Setting this in the denominator,

$$P(y=+1|x) = \frac{P(x|y=+1) \times P(y=+1)}{\sum_{z \in \{-1,1\}} P(x|y=z) \times P(y=z)}$$

Expanding the sum,

$$P(y=+1|x) = \frac{P(x|y=+1) \times P(y=+1)}{P(x|y=-1) \times P(y=-1) + P(x|y=+1) \times P(y=+1)}$$

Knowing,

$$P(y=+1) = p$$

Then,

$$1 - P(y=+1) = 1 - p = P(y=-1)$$

Rewrite,

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

Set the numerator to 1,

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

Using rules of log,

$$e^{\log C} \equiv \exp(\log C) = C, C \in R$$

Thus,

$$P(y=+1|x) = \frac{1}{1 + \exp(\log \frac{P(x|y=-1) \times (1-p)}{P(x|y=+1) \times p})}$$

This is in form,

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

Where A, B are constants

$$\boxed{A = P(x|y=-1) \times (1-p)}$$

$$\boxed{B = P(x|y=+1) \times p}$$

2. Rewriting,

$$P(X|y=+1) = P(X_1 = x_1, X_2 = x_2, ...X_d = x_d | y=+1)$$

We have a similar expression with $y = -1$. We assume that, $X_1, X_2, ...X_d$ are i.i.d and that they are statistically independent,

$$P(X_1 = x_1, ...X_d = x_d | y=+1) = \prod_{j=1}^{d} P(X_j = x_j | y=+1)$$

Similar for $y = -1$,

$$P(X_1 = x_1, ...X_d = x_d | y=-1) = \prod_{j=1}^{d} P(X_j = x_j | y=-1)$$

Rewrite log expression using the normal distribution provided in the question,

$$\log(\frac{A}{B}) = \log(\frac{1 - p \times N(\mu_-, I)}{p \times N(\mu_+, I)})$$

Separate the constant from RHS and rewrite using conditioning,

$$\log(\frac{1-p}{p}) + \log\frac{\prod_{j=1}^{d} P(X_j = x_j|y=-1)}{\prod_{j=1}^{d} P(X_j = x_j|y=+1)}$$

Holding off constant term till the end, we rewrite log as,

$$\log\frac{\prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_j - \mu_{-,j})^2}{2}\right)}{\prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_j - \mu_{+,j})^2}{2}\right)}$$

Rewrite exponent power as a summation,

$$\log\frac{\left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left(\sum_{j=1}^{d} -\frac{(x_j - \mu_{-,j})^2}{2}\right)}{\left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left(\sum_{j=1}^{d} -\frac{(x_j - \mu_{+,j})^2}{2}\right)}$$

Cancelling the common term in the log and focusing on the power of the exponential,

$$\sum_{j=1}^{d} -\frac{(x_j - \mu_{-,j})^2}{2} - \sum_{j=1}^{d} -\frac{(x_j - \mu_{+,j})^2}{2}$$

Rewriting and expanding,

$$\sum_{j=1}^{d} \frac{(x_j^2 - 2x_j\mu_{+,j} + \mu_{+,j}^2)}{2} - \sum_{j=1}^{d} \frac{(x_j^2 - 2x_j\mu_{-,j} + \mu_{-,j}^2)}{2}$$

Simplifying,

$$\sum_{j=1}^{d} (x_j(\mu_{-,j} - \mu_{+,j}) + \frac{1}{2}(\mu_{+,j}^2 - \mu_{-,j}^2))$$

We know $\mu_+, \mu_- \in R^d$,

$$(\mu_- - \mu_+)^T x + \frac{1}{2}(\mu_+^T\mu_+ - \mu_-^T\mu_-)$$

The complete expression for $\log\frac{A}{B}$ is,

$$\boxed{\log\frac{A}{B} = (\mu_- - \mu_+)^T x + \frac{1}{2}(\mu_+^T\mu_+ - \mu_-^T\mu_-) + \log(\frac{1-p}{p})}$$

Where the constants are,

$$\boxed{w = \mu_- - \mu_+}$$

$$\boxed{b = \frac{1}{2}(\mu_+^T\mu_+ - \mu_-^T\mu_-) + \log(\frac{1-p}{p})}$$

3. $P(y|x)$ is the conditional probability of the random variable y given a value of x. From part (i),

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

Similar expression for $y = -1$ can be derived,

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

Using log-properties and rewriting,

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

We can add a y term too,

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

Replace $\log\frac{A}{B}$ in terms of w, x, b

$$\boxed{P(y|x) = \frac{1}{1+\exp(y \times (w^T x + b))}}$$

5 Linear regression: 14pts + 1pt

3. Code for the `plot_linear()` function is provided in the `hw1.py` file. See the figure below for an example plot.

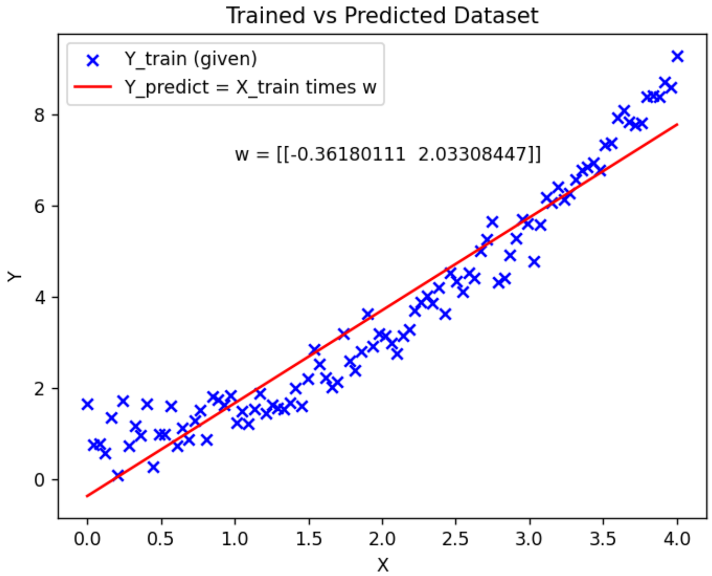

Figure 1: Example plot generated using `plot_linear()`.