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.

Reminder: Answers must be typeset. LaTeXand other methods of typesetting math are accepted.

1 Short answer: 10pts

Each question is worth 2 points. One-sentence explanations are allowed but not necessary for full credit.

1. A (1-)nearest neighbour model is trained on a dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)}\}_{i=1}^{N}$ where $x^{(i)} \in \mathbf{R}^d \ \forall i$. That is, there are N training images, each of which is dimension d. It is then used to evaluate M test images. What is the time complexity of the test-time evaluation? Use big-O notation.

   Answer: $O(dNM)$. $O(dM \log N)$ can also be achieved by sorting the training points with a K-D tree! Both answers will receive full credit.

2. Consider two different k-nearest neighbor models: one has $k = 1$ and one with $k = 10$. In general, which would you expect to expect to have a "smoother" decision boundary?

   Answer: The larger, $k = 10$.

3. Let g be the logical OR function, defined on the feature space $\{+1, -1\}^2$, which maps $g(+1, +1) = +1, g(-1, +1) = +1, g(+1, -1) = +1, g(-1, -1) = -1$. Given a linear classifier $h(x) = \text{sign}(\boldsymbol{w} \cdot \boldsymbol{x} + b)$, give a valid $(\boldsymbol{w}, b)$ pair that matches ground truth g. Let $\text{sign}(z) = +1$ for $z \geq 0$ and -1 otherwise.

   Answer: One set of valid answers is $\boldsymbol{w} = (1, 1), b \in [0, 2)$

4. For real matrix $A \in \mathbb{R}^{n \times m}$, what relationship does the largest singular value of A have with the largest eigenvalue of $A^\top A$?

   Answer: $\sigma_{\max}(A) = \sqrt{\lambda_{\max}(A^\top A)}$

5. As mentioned in lecture, image data does not normally satisfy the Naive Bayes assumption. Give one additional example of a real-world situation in which the Naive Bayes assumption is violated.

   Answer: Most real-world examples should work. Something like tabular healthcare data, where height/weight/sex might be correlated.

2 Linear Regression: 10pts

Consider a data matrix $X \in \mathbb{R}^{n \times d}$ with rows $(\boldsymbol{x}_i)_{i=1}^n$. Assume that $d > n$ and that X is full-rank; that is, $\text{rank}(X) = n$.

1. (5pts) Show that there exists a $\boldsymbol{w}$ such that the empirical risk with squared loss is zero, i.e., that $X\boldsymbol{w} = \boldsymbol{y}$.

   Answer: 2 of many possible answers below:

   - X must have n independent columns, so $\boldsymbol{y}$ must be the image/column span of X. This means there is some linear combination of the columns of X that gets $\boldsymbol{y}$, which defines solution $\boldsymbol{w}$.
   - This is the same as solving a linear system of n equations with d variables. For this answer to get full credit, there needs to be some discussion about why this isn't a degenerate system.

2. (2pts) Let the SVD of X be $X = U\Sigma V^\top$. What is the rank of Σ?

   Answer: 2 of many possible answers below:

   - $\text{rank}(U\Sigma V^\top) = \text{rank}(\Sigma)$ because U, V are invertible. Proof: $\text{rank}(AU) \leq \text{rank}(A)$ and $\text{rank}(A) \leq \text{rank}(AUU^{-1}) \leq \text{rank}(AU)$. Then, multiplying a matrix by an invertible square matrix does not change its rank. This applies for U, V square and $\Sigma \in \mathbb{R}^{n \times d}$.
   - By the definition of SVD, Σ is a diagonal matrix with n nonzero entries. Thus, $\text{rank}(\Sigma) = n$.

3. (3pts) Show that $XX^\top$ is invertible.

   Answer: $XX^\top = U\Sigma V^\top V \Sigma U^\top = U\Sigma^2 U^\top$. Note that $U \in \mathbb{R}^{n \times n}$ for SVD (or $n \times r$ for rank r in a compact SVD, but $r = n$ in this case), and Σ^2 is also $n \times n$ ($r \times r$). U is orthogonal and Σ is full rank, so $XX^\top$ is full-rank and therefore invertible.

3 SVM: 10 pts

1. (2pts) Recall the dual of hard-margin SVM for binary classification:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j : \boldsymbol{\alpha} \in \mathbb{R}^n, \boldsymbol{\alpha} \geq 0$$

   What is the smallest number of support vectors for a d-dimensional dataset $\mathcal{D} =$

$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{10,000}$? In other words, $\boldsymbol{x}_i \in \mathbb{R}^d \, \forall i \in [n]$. Assume that $\mathcal{D}$ is linearly separable and that there exists at least one point in each class.

Answer: 2, the number of classes. They must be equidistant from the hyperplane $\boldsymbol{w}$, and they define the margin.

2. (3pts) Recall from lecture that for any datapoint $(\boldsymbol{x}_i, y_i)$, if $\alpha_i > 0$, then $\boldsymbol{x}_i$ is a support vector. We will complete the definition of support vector to be *any point that lies on the margin*; in other words, a given point $(\boldsymbol{x}_i, y_i)$ is a support vector if and only if $y_i(\boldsymbol{w}^*)^\top \boldsymbol{x}_i = 1$.

   Now, let an optimal solution to the dual on our dataset $\mathcal{D}$ be $\boldsymbol{\alpha} = [10, 2, 3, 0, \cdots, 0]$ (omitted elements are all 0). What is the smallest possible and largest possible number of support vectors in $\mathcal{D}$?

   Hint: A solution to the dual might not be unique. Is it possible for a point to contribute to the optimal $\boldsymbol{w}^*$ for one solution $\boldsymbol{\alpha}$, but have 0 contribution/weight in a different solution $\boldsymbol{\alpha}'$? Is this point a support vector?

   Answer: Minimum is 3, as any entry of $\alpha_i > 0$ indicates a support vector where $y_i \boldsymbol{x}_i^\top \boldsymbol{w} = 1 \; \forall i \in [n]$.

   Maximum is 10,000 (also accept n or "number of datapoints"). Visualize for example, 5000 positive points that are all the same, stacked on top of each other, and 5000 negative points all the same. Or, another example might be if 10,000 points are all γ distance away from the decision boundary, running in two parallel lines around the decision boundary.

   This is a geometric interpretation of a single primal solution paired with non-unique dual solutions and can result from a noninvertible kernel.

3. Recall the XOR problem. In this question, we want to model the function $g_{XNOR} : \{-1, 1\}^2 \to \{-1, 1\}$:

$$g_{XNOR}(-1, -1) = g_{XNOR}(1, 1) = 1$$
$$g_{XNOR}(1, -1) = g_{XNOR}(-1, 1) = -1$$

   To solve this problem, we need a nonlinear mapping. Consider the following kernel:

$$k(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^\top \boldsymbol{z} + 1)^2$$

   (a) (3pts) Write out a feature mapping $\phi : \mathbb{R}^2 \to \mathbb{R}^d$ that induces this kernel. In other words, what is one ϕ that satisfies $\phi(\boldsymbol{x})^\top \phi(\boldsymbol{z}) = k(\boldsymbol{x}, \boldsymbol{z})$?

   Answer: $\phi(\boldsymbol{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1)^\top$

(b) (2pts) Find a solution $\boldsymbol{w}$ such that $h(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^\top \phi(\boldsymbol{x})) = g_{XNOR}(\boldsymbol{x})$

Answer: $\boldsymbol{w} = (0, 0, 1, 0, 0, 0)^\top$

4 Gaussian Naive Bayes: 15pts

Recall that the Bayes Classifier is

$$h(\boldsymbol{x}) = \underset{y}{\text{argmax}}\, P(y|\boldsymbol{x})$$

We will work with binary classification: $y_i \in \{-1, +1\}\, \forall i \in [n]$. The feature vectors are now continuous: $\boldsymbol{x} \in \mathbb{R}^d$.

1. (5pts) Assume that we have a prior $P(y = +1) = p$ for some $p \in (0, 1)$.

   Show that the predictor $P(y = +1|\boldsymbol{x})$ can be written can be written $\frac{1}{1+\exp\left(\log \frac{A}{B}\right)}$

   where A, B are expressions in terms of $p, P(\boldsymbol{x}|y = +1), P(\boldsymbol{x}|y = -1)$.

   Answer:

$$\begin{aligned} P(y = +1|\boldsymbol{x}) &= \frac{p \cdot P(\boldsymbol{x}|y = +1)}{p \cdot P(\boldsymbol{x}|y = +1) + (1 - p) \cdot P(\boldsymbol{x}|y = -1)} \\ &= \frac{1}{1 + \frac{(1-p)P(\boldsymbol{x}|y=-1)}{pP(\boldsymbol{x}|y=+1)}} \\ &= \frac{1}{1 + \exp\left(\log \frac{(1-p)P(\boldsymbol{x}|y=-1)}{pP(\boldsymbol{x}|y=+1)}\right)} \end{aligned}$$

2. (8pts) Consider a Gaussian Naive Bayes model. Let x_j be the jth element of $\boldsymbol{x}$. Let the data be generated as follows for $\boldsymbol{\mu}_+, \boldsymbol{\mu}_- \in \mathbb{R}^d$ and $I \in \mathbb{R}^{d\times d}$ the identity matrix:

$$P(\boldsymbol{x}|y = +1) = \mathcal{N}(\boldsymbol{\mu}_+, I),\ P(\boldsymbol{x}|y = -1) = \mathcal{N}(\boldsymbol{\mu}_-, I)$$

   For example, the positive class has distribution

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

   Show that the expression from the previous part, $\log \frac{A}{B}$, can be written in the form $\boldsymbol{w}^\top \boldsymbol{x} + b$, where $\boldsymbol{w}$ and b are expressions in terms of $p, \boldsymbol{\mu}_+, \boldsymbol{\mu}_-$. Identify assumptions and definitions used in your derivation.

Answer:

$$\begin{aligned}
\log \frac{(1-p)P(\boldsymbol{x}|y=-1)}{pP(\boldsymbol{x}|y=+1)} &= \log \frac{1-p}{p} + \log \frac{P(\boldsymbol{x}|y=-1)}{P(\boldsymbol{x}|y=+1)} \\
&= \log \frac{1-p}{p} + \log \frac{\prod_{j=1}^{d} P(x_j|y=-1)}{\prod_{j=1}^{d} P(x_j|y=+1)} \\
&= \log \frac{1-p}{p} + \sum_{j=1}^{d} \log \frac{P(x_j|y=-1)}{P(x_j|y=+1)} \\
&= \log \frac{1-p}{p} + \sum_{j=1}^{d} -1/2 * ((x_j - \mu_{-,j})^2 - (x_j - \mu_{+,j})^2) \\
&= \log \frac{1-p}{p} + \sum_{j=1}^{d} \left((\mu_{-,j} - \mu_{+,j}) x_j + \frac{\mu_{+,j}^2 - \mu_{-,j}^2}{2} \right) \\
&= (\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)^\top \boldsymbol{x} + \log \frac{1-p}{p} + \frac{\|\boldsymbol{\mu}_+\|^2 - \|\boldsymbol{\mu}_+\|^2}{2} \\
&= \boldsymbol{w}^\top \boldsymbol{x} + b
\end{aligned}$$

3. (2pts) Write a single expression for $P(y|\boldsymbol{x})$ as a function of $y, \boldsymbol{x}, \boldsymbol{w}, b$.

   Answer: For the above result where $\boldsymbol{w} = \boldsymbol{\mu}_- - \boldsymbol{\mu}_+$ and $b = \log \frac{1-p}{p} + \frac{\|\boldsymbol{\mu}_+\|^2 - \|\boldsymbol{\mu}_-\|^2}{2}$:

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

   Alternatively, we can see that Gaussian NB and logistic regression produce the same model under the NB assumption. Flip the signs, and let $\boldsymbol{w} = \boldsymbol{\mu}_+ - \boldsymbol{\mu}_-$ and $b = \log \frac{p}{1-p} + \frac{\|\boldsymbol{\mu}_-\|^2 - \|\boldsymbol{\mu}_+\|^2}{2}$ to get:

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

5 Linear regression: 14pts + 1pt

Recall that the empirical risk in the linear regression method is defined as

$$\mathcal{R}(w) := \frac{1}{2n} \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

where $x_i \in \mathbb{R}^d$ is a data point and y_i is an associated label.

1. (10.5pts) **Implement** the linear regression method using gradient descent in `linear_gd(X, Y, lrate, num_iter)` function in `hw1.py`. You are given a training set `X` as input and training labels `Y` as input along with a learning rate `lrate` and maximum number of iterations `num_iter`. Using gradient descent find parameters w that minimize the empirical risk $\mathcal{R}(w)$. One iteration is equivalent to one full data gradient update step. Use a learning rate of `lrate` and only run for `num_iter` iterations. Use $w = 0$ as your initial parameters, and return your parameters w as output.
2. (3.5pts) **Implement** linear regression by setting the gradient to zero and solving for the variables, in `linear_normal(X,Y)` function in `hw1.py`. You are given a training set `X` as input and training labels `Y` as input. Return your parameters w as output.
3. (1pt) Implement the `plot_linear()` function in `hw1.py`. Use the provided function `utils.load_reg_data()` to generate a training set `X` and training labels `Y`. Plot the curve generated by `linear_normal()` along with the points from the data set. Return the plot as output. **Include the plot in your written submission.**