**CS461 Midterm (March. 13, 2025)**

| | |
|---|---|
| CS461 Section #: | |
| Name: | Solution |
| NetID: | Total: 130. |

**0. (3p/each) True/ False Questions**

**Part1:**

- The posterior probability $P[A|B]$ is sensitive to the prior probability $P[A]$. (True / False)
- Maximum likelihood estimation is sensitive to the data size. (True / False)
- In an over-determined system $Ax = y$, if $y$ is on the column space of $A$, a solution always exists. (True / False)
- By increasing the amount of training data, we can reduce the bias of a model. (True / False)
- Ridge regularization (adding $\lambda||w||^2$ to an objective) can be interpreted as solving an optimization problem with inequality constraints. (True / False)

**Part2**

- Logistic regression algorithm does not converge if data is not linearly separable. (True/ False)
- MAP classification minimizes the error rate. (True/False)
- Both Naive Bayes and logistic regression models learn posterior density $P(C_k|x)$. (True / False)
- In perceptron, when data is linearly separable, depending on the initial point, the convergence point will be different. (True / False)
- Even with an imbalance between positive and negative samples, accuracy is not affected, as it accounts for both false negatives and false positives. (True/ False)

**Part3**

- In general, SVM exhibits higher sensitivity (or high variance) to different training data compared to other algorithms as kernel methods like Gaussian map data into an infinite-dimensional feature space to construct a maximal margin classifier. (True/ False.)
- In SVM, removing non-support vectors does not change the decision boundary. (True /False)
- When data is not linearly separable then no way for a hard margin classifier to converge. (True /False)
- When data is linearly separable, soft and hard margin SVM will result in the same classifier. (True / False)
- A soft-margin SVM generally results in a larger margin than a hard-margin SVM. (True / False)

**2.1(10p)** [Bayes Rule] Before going on vacation, you ask your friend to water your ailing plant. Without water, the plant has an $A$ percent chance of dying. Even with proper watering, it has a $B$ percent chance of dying. And the probability that your friend will forget to water it is $C$ percent. If the plant survived in the week, what is the probability that your friend watered it? (write your solution using $A$, $B$, $C$.)

· Sol) $P(w|s) = \dfrac{P(s|w)\cdot P(w)}{P(s|w)\cdot P(w) + P(s|w^c)\cdot P(w^c)} = \dfrac{(1-B)\cdot(1-C)}{(1-B)(1-C)+(1-A)\cdot C}$

· Rub)

· only Bayes Rule : 5 - P

· only $P(w|s)$ : 3 - P

**2.2(10p)** [KKT conditions] Compute the optimal solution $x_1^*, x_2^*, x_3^*$ for the problem below. Please show all intermediate steps clearly.

$$\begin{aligned} \min \quad & \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) \\ \text{subject to} \quad & x_1 + x_2 + x_3 \le -3 \end{aligned}$$

· Sol) $\mathcal{L}(x_1, x_2, x_3, \lambda) = \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) + \lambda(x_1 + x_2 + x_3 + 3)$

① · $\nabla_x \mathcal{L}(x_1^*, x_2^*, x_3^*, \lambda^*) = 0 \iff \begin{bmatrix} x_1^* \\ x_2^* \\ x_3^* \end{bmatrix} + \lambda^* \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

$\iff x_1^* = x_2^* = x_3^* = -\lambda^*$

② $\begin{cases} \lambda^* = 0 \;\&\; x_1^* + x_2^* + x_3^* < -3 \Rightarrow & 0 < -3 \quad \text{Contradiction} \\ \lambda^* > 0 \;\&\; x_1^* + x_2^* + x_3^* = -3 \Rightarrow & -\lambda^* = -3 \\ & \lambda^* = 1 \end{cases}$

· Rub
- missing complementary slackness — 0P
- right up to ① — 5P
- only answer is right — 3P

$\therefore x_1^* = x_2^* = x_3^* = -1.$

3 [Linear Regression] Suppose you are given a set of data $\{(x_i, y_i) \mid i = 1, 2, \ldots, n, x_i \in \mathcal{R}, y_i \in \mathcal{R}\}$. You know that the data is artificially generated by a polynomial $f(x)$ but don't know the exact degree of it. The data contains an additive noise $\epsilon$ and it follows Gaussian.

$$y_i = f(x_i) + \epsilon_i \mid i = 1, 2, \ldots, n,$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2 = 1.0e^{-3}) \quad \forall i$$

3.1(5p) Suppose you trained a MMSE linear regression using the basis functions of $[1, x]$ and $N$ data ponts. The Mean Square Error (MSE) for train and test are shown below. What is a possible reason for the large MSE values?

| | |
|---|---|
| train | MSE$\gg 1.0e-3$ |
| test | MSE$\gg 1.0e-3$ |

- the basis function of [1, x] is not sufficient to capture the data complexity.
- it is hard to tell whether the data points are enough or not.

3.2(5p) Suppose you trained a MMSE linear regression using the basis functions of $[1, x, x^2, x^3, x^4]$ and the same number of data points $N$ in problem 3.1. The Mean Square Error (MSE) for train and test are shown below. Based on values, what issue can you identify?

| | |
|---|---|
| train | $\approx 0$ |
| test | MSE$\gg 1.0e-3$ |

- The model shows a large variance.
- The model is complex. but the number of data points is not enough to properly constrain it.

3.3(5p) What will be two possible solutions to address the issue in problem 3.2?

- Regularization or Reduce the complexity
- increase # data points

3.4(5p) Suppose data is generated without intrinsic error $\epsilon$ and the polynomial is $f(x) = 1 + x + x^2 + x^3 + x^4 + x^5$. Then, how many data points will be sufficient to recover the original polynomial using MMSE linear regression? Please provide a brief explanation for your reasoning.

Since there are 6 unknown parameters, six different data points will be sufficient to recover the original polynomial.

**4. [Linear Classification]** Suppose you train a Gaussian discriminant classifier based on the six data points below. When assuming equal variances for the two classes, the binary decision rule of GDA is formulated as follows.

$$P[x|C_+] \cdot P[C_+] \underset{C_+}{\overset{C_-}{\lessgtr}} P[x|C_-] \cdot P[C_-]$$

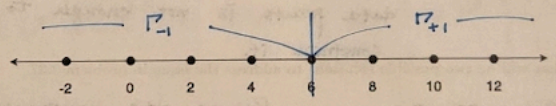
$$\leftrightarrow -\frac{1}{2}(x-\mu_+)^2 \underset{C_+}{\overset{C_-}{\lessgtr}} -\frac{1}{2}(x-\mu_-)^2, \text{ where } P[C_+] = P[C_-] = \frac{1}{2} \text{ and } \sigma_+ = \sigma_- = \sigma$$

$$\leftrightarrow x \underset{C_+}{\overset{C_-}{\lessgtr}} \frac{1}{2} \cdot (\mu_- + \mu_+)$$

**4.1(5p) Given the six data points below, is the data linearly separable?** Yes

| data num | $x$ | class ($t$) |
|---|---|---|
| $d_1$ | $-10$ | $-1$ |
| $d_2$ | $0$ | $-1$ |
| $d_3$ | $+10$ | $-1$ |
| $d_4$ | $+11$ | $+1$ |
| $d_5$ | $+12$ | $+1$ |
| $d_6$ | $+13$ | $+1$ |

**4.2(5p) Compute a decision boundary and region. Please draw them below based on the MAP rule provided above.**




**4.3(5p) The decision boundary in problem 4.2 separates the train data? What is the margin of the classifier? hint: the margin is the smallest value of $t_i(w \cdot x_i + b)$ for data $d_i$.**

$d_1$ : $-1 \cdot (-10 - 6) = 16$

$d_2$ : $-1 \quad (0 - 6) = 6$

$d_3$ : $-1 \quad (10 - 6) = -4$ (Rubric : 4 is okay. too).

$d_4$ : $+1 \quad (11 - 6) = 5$

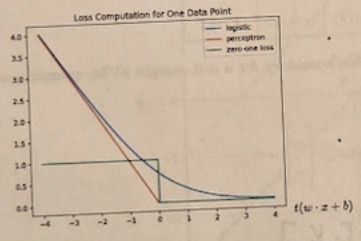
$d_5$ : $+1 \quad (12 - 6) = 6$

$d_6$ : $+1 \quad (13 - 6) = 7$

4.4(5p) Suppose perceptron and logistic regression are trained based on the same data above. Do you think the classifiers linearly separate the training data?

| | separation | no separation |
|---|---|---|
| Perceptron | ✓ | |
| Logistic Regression | ✓ | |

4.5(5p) In class, we studied the loss of a single data point for perceptron and logistic regression, as shown in the figure below. Based on the margin in problem 4.3 and the figure, which classifier will yield the largest margin among perceptron, logistic regression, and GDA? And, which classifier has the second largest? Please provide a brief explanation for your reasoning.




- logistic regression > perceptron > GDA.
- Rubric: . No point when the order is wrong
  - 3-p: Not provided explanation or wrong explanation.