A2 (Written Piece)

Q1) Consider an ensemble binary classification technique that is based on N IID classifiers $\{C_i\}_{i=1}^{N}$, where N is odd. For any given dp, each base classifier C_i predicts the class label and the most commonly predicted class label and the most commonly predicted class label across all base classifiers is selected as the final prediction. Assuming that all base classifiers have an error rate ε (i.e., the prob of making an incorrect prediction), express the error rate of the ensemble technique in terms of ε and N, and explain your answer.

Ans) Let X be the prob dist representing the error rate of the ensemble technique.

In order for the ensemble technique to make an incorrect prediction, more than half (i.e., at least $\lceil \frac{N}{2} \rceil$) must make an incorrect prediction.

Consider the case where exactly x C_i's are erroneous. Then, the probability is given by:

$$P(X=x) = \binom{N}{x} (\varepsilon)^x (1-\varepsilon)^{N-x}, \text{ for } x = \lceil \tfrac{N}{2} \rceil, \ldots, N.$$

From this example, we see that $X \sim \text{Binomial}(N, \varepsilon)$.
Therefore, to obtain the probability of <u>at least</u> $\lceil \frac{N}{2} \rceil$ C_i's making an incorrect prediction, we simply take the sum of the probabilities:

$$P(X \geq x = \lceil \tfrac{N}{2} \rceil) = \sum_{x=\lceil \frac{N}{2} \rceil}^{N} \binom{N}{x} (\varepsilon)^x (1-\varepsilon)^{N-x}.$$

Q2) Consider a binary classification scenario w/ 1-dim data and GCC densities. That is, $P(x|c_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(\frac{-(x-\mu_i)^2}{2\sigma_i^2}\right)$, where $i \in \{1,2\}$.

Prove that, if $\sigma_1 = \sigma_2$, then the decision boundary is linear.

pf.) Let $P(x|c_i) = G(x;\mu_i,\sigma_i)$. Suppose $\sigma_1 = \sigma_2$. Let $a(x)$ be the decision boundary of the GCC.

$$a(x) = -\frac{1}{2}(x-\mu_1)^T\sigma_1^{-1}(x-\mu_1) - \cancel{\frac{1}{2}\ln(\sigma_1)} + \frac{1}{2}(x-\mu_2)^T\sigma_2^{-1}(x-\mu_2) + \cancel{\frac{1}{2}\ln(\sigma_2)}$$

$$= \frac{\sigma^{-1}}{2}\left[(x-\mu_2)^T(x-\mu_2) - (x-\mu_1)^T(x-\mu_1)\right]$$

$$= \frac{\sigma^{-1}}{2}\left[(x-\mu_2)^2 - (x-\mu_1)^2\right]$$, x is 1-dim $\Rightarrow \mu_i$ is 1-dim $\Rightarrow (x-\mu_i)^T = (x-\mu_i)$

$$= \frac{\sigma^{-1}}{2}\left[\cancel{x^2} - 2x\mu_2 + \mu_2^2 - \cancel{x^2} + 2x\mu_1 - \mu_1^2\right]$$

$$= x\sigma^{-1}(\mu_1-\mu_2) + \frac{\sigma^{-1}(\mu_2^2-\mu_1^2)}{2}$$. Let $a = \sigma^{-1}(\mu_1-\mu_2) \in \mathbb{R}$, $b = \frac{\sigma^{-1}(\mu_2^2-\mu_1^2)}{2} \in \mathbb{R}$

$$= ax + b.$$

$\therefore$ The decision boundary is linear, as wanted. ■

Q3) Assume we want to build a logistic reg model to classify fruit as orange/non-orange using it's width and height. The training data to be used is as follows:

Width	Height	Orange
4	4	Yes
6	4	Yes
6	5	Yes
6	8	No
6	10	No
8	8	Yes
8	10	No

Look at $\frac{\partial}{\partial w} L(w)$ (W8)

a) Write the corresponding opt prob in terms of the data provided above and specify the params to be estimated.

(Opt Prob) Minimize $L(w) = \sum_{i=1}^{7} y_i \log P(\theta|x_i) + (1-y_i)\log(1-P(\theta|x_i))$,

where $P(\theta|x_i) = \frac{1}{1+e^{-w^T+b}} = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$.

Params to estimate: w_2, w_1, w_0.

Constraints: None.

b) Perform 3 iters of the GD alg to determine the params assuming that the step size (λ) is 0.01 and the init estimate is [0.3, -0.2, 0.7] (note that 0.7 corresponds to the bias.) For each estimate, including the initial one, you are required to report the following:

- The val of the estimate.
- The accuracy of the resulting logistic reg model when applied to the training data.

Note that you do not need to the computations manually. You might want to use a spreadsheet or write a simple prog to do that.

Ans)

Iter	Val	Accuracy
0	[0.3 -0.2 0.7]	57.1%
1	[0.20, -0.35, 0.69]	85.7%
2	[0.28, -0.31, 0.70]	100.0%
3	[0.27, -0.36, 0.70]	100.0%

Wow! 100% accuracy!

c) Classify the following dps using the model you obtained in (b): (3,3), (4,10), (9,8), (9,10).

Ans) Obj function: $p(c_i|x) = \frac{1}{1+e^{-(0.27x_1-0.36x_2+0.70)}}$

$p(c_i|(3,3)) \approx 0.61 > 0.5 \implies (3,3) \in c_1 = \text{Orange.}$

$p(c_i|(4,10)) \approx 0.14 < 0.5 \implies (4,10) \in c_2 = \text{Non-Orange.}$

$p(c_i|(9,8)) \approx 0.56 > 0.5 \implies (9,8) \in c_1 = \text{Orange.}$

$p(c_i|(9,10)) \approx 0.38 < 0.5 \implies (9,10) \in c_2 = \text{Non-Orange}$

An even split! Man, 100% (training) accuracy models feel nice...

d) An advantage of logistic reg is that it doesn't require us to compute the covariance matrix (unlike, say, GCCs) and is thus more computationally efficient than other methods that require it, as it is typically the most expensive operation in building a model.

A disadvantage of logistic reg is that it doesn't have a closed-form solution, and must therefore rely on local search algorithms such as gradient descent to obtain its params. This method doesn't guarantee the optimal params as local search algs may get stuck w/in local minima.