Problem 1 *[20 points]*

Preliminaries: Suppose $X = \{X_i\}_{i=1}^D \in \mathbb{R}^D$ represents the features and $Y \in \{0, 1\}$ represents the class labels. Let the following assumptions hold:

1. The label variable Y follows a Bernoulli distribution, with parameter $\boldsymbol{\pi} = P(Y = 1)$.
2. For each feature X_j, we have $P(X_j|Y = y_k)$ which follows a Gaussian distribution $N(\mu_{jk}, \sigma_j)$.

Using the Naive Bayes assumption, *"for all $j' \neq j$, X_j and $X_{j'}$ are conditionally independent given Y"*, compute $P(Y = 1|X)$ and show that it can be written in the following form:

$$P(Y = 1|X) = \frac{1}{1 + \exp(-w_0 + \boldsymbol{w}^T \boldsymbol{X})}$$

Specifically, find the explicit form of w_0 and $\boldsymbol{w}$ in terms of $\boldsymbol{\pi}, \mu_{jk}$, and σ_j, for $j = 1, \ldots, D$, and $k \in \{0, 1\}$.
Solution: For ease of indexing, and without loss of generality, let $y_0 = 0$ and $y_1 = 1$:

$$P(Y = 1|X) = P(X|Y = 1)\frac{P(Y = 1)}{P(X)} \ = \pi \frac{P(X|Y = 1)}{P(X)}.$$

with

$$P(Y = 1) = \pi$$

Now

$$P(X) = \sum_y P(X|Y = y) = \pi P(X|Y = 1) + (1 - \pi)P(X|Y = 0)$$

Thus

$$P(Y = 1|X) = \frac{1}{1 + \frac{1-\pi}{\pi}\frac{P(X|Y=0)}{P(X|Y=1)}}$$

Explicitly, using

$$P(X|Y = y_k) = \prod_{j=1}^{D} (2\pi\sigma_j^2)^{-\frac{1}{2}} \exp(-(2\sigma_j^2)^{-1}(x_j - \mu_{jk})^2)$$

We get

$$P(Y = 1|X) = \frac{1}{1 + (\frac{1}{\boldsymbol{\pi}} - 1)\prod_{j=1}^{D} \exp((2\sigma_j^2)^{-1}((x_j - \mu_{j1})^2 - (x_j - \mu_{j0})^2)}$$

Consider only the second term of the denominator:

$$\begin{aligned}
&(\frac{1}{\boldsymbol{\pi}} - 1)\prod_{j=1}^{D} \exp((2\sigma_j^2)^{-1}((x_j - \mu_{j1})^2 - (x_j - \mu_{j0})^2) \\
&\qquad = \exp(\log(\frac{1}{\boldsymbol{\pi}} - 1)) \exp(\sum_j (2\sigma_j^2)^{-1}(x_j^2 - 2x_j\mu_{j1} + \mu_{j1}^2 - x_j^2 + 2x_j\mu_{j0} - \mu_{j2}^2)) \\
&\qquad = \exp(\log(\frac{1}{\boldsymbol{\pi}} - 1) + \sum_j (2\sigma_j^2)^{-1}(\mu_{j1}^2 - \mu_{j2}^2) + \sum_j (\sigma_j^2)^{-1}(\mu_{j0} - \mu_{j1})x_j)
\end{aligned}$$

Comparing this with the second term in the denominator of

$$P(Y = 1|X) = \frac{1}{1 + \exp(-w_0 + \boldsymbol{w}^T \boldsymbol{X})}$$

gives us:

1. $w_0 = -\log(\frac{1}{\boldsymbol{\pi}} - 1) - \sum_j \left[(2\sigma^2)^{-1}(\mu_{j1}^2 + \mu_{j2}^2)\right]$.

2. $w_j = \frac{\mu_{j0} - \mu_{j1}}{\sigma_j^2}, for j >= 1$.

1.2) Suppose we are given a training set with N examples $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}, y_2), \ldots, (\boldsymbol{x}_N, y_N)$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iD}^T$ is a $D - dimensional$ feature vector, and $y_i \in \{0, 1\}$ is its corresponding label. Using the assumptions in 1.1, provide the maximum likelihood estimation for the parameters of the Naive Bayes with Gaussian assumption.

Solution:

$$LL = \log \prod_{c=0}^{1} \prod_{\substack{i=1 \\ y_i=c}}^{N} \prod_{j=1}^{D} \left[(2\pi\sigma_j^2)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma_j^2}(x_{ji} - \mu_{jc})^2 \right\} \right] = -\sum_{c=0}^{1} \sum_{\substack{i=1 \\ y_i=c}}^{N} \sum_{j=1}^{D} \left[\frac{1}{2}\log(2\pi\sigma_j^2) - \frac{1}{2\sigma_j^2}(x_{ji} - \mu_{jc})^2 \right]$$

$$\frac{\partial}{\partial \mu_{jc}} LL = \sum_{\substack{i=1 \\ y_i=c}}^{N} \frac{x_{ji} - \mu_{jc}}{\sigma_j^2} = 0$$

Set this equal to zero and split the summation.

$$\sum_{\substack{i=1 \\ y_i=c}}^{N} x_{ji} = \sum_{\substack{i=1 \\ y_i=c}}^{N} \mu_{jc} = N_c \mu_{jc}$$

$$\mu_{jc} = \frac{1}{N_c} \sum_{\substack{i=1 \\ y_i=c}}^{N} x_{ji}$$

For σ_j:

$$\frac{\partial}{\partial(\sigma_j^2)} LL = \sum_{c=0}^{1} \left[-\sum_{\substack{i=1 \\ y_i=c}}^{N} \frac{1}{2}\frac{1}{\sigma_j^2} + \sum_{\substack{i=1 \\ y_i=c}}^{N} \frac{1}{2\sigma_j^4}(x_{ji} - \mu_{jc})^2 \right]$$

Set this equal to zero and separate:

$$\sum_{c=0}^{1} \sum_{\substack{i=1 \\ y_i=c}}^{N} \frac{1}{\sigma_j^2} = \sum_{c=0}^{1} \sum_{\substack{i=1 \\ y_i=c}}^{N} \frac{1}{\sigma_j^4}(x_{ji} - \mu_{jc})^2$$

$$\sigma_j^2 = \frac{1}{N} \sum_{c=0}^{1} \sum_{\substack{i=1 \\ y_i=c}}^{N} (x_{ji} - \mu_{jc})^2$$

$$\sigma_j = \sqrt{\frac{1}{N} \sum_{c=0}^{1} \sum_{\substack{i=1 \\ y_i=c}}^{N} (x_{ji} - \mu_{jc})^2}$$

Problem 2 *[10 points]*

Part (a): **Solution:**

1. Instantiate a mesh-grid G with resolution M on the 3-cube. That is, build the set of all 3-tuples (x, y, z) where the possible values of each element are in the set $\{0, \frac{1}{M}, \frac{2}{M}, \ldots, \frac{M-1}{M}, 1\}$.
2. For each mesh-point $p_i \in G$ the mesh-grid, calculate the leave-one-out training accuracy a_i.
3. Whichever a_i is highest, choose the corresponding p_i as optimal weight.

Part (b): **Solution:** This approach does not work in practice; while for finite D dimension we can still enumerate and process all of the test points in theory, in practice this grows exponentially fast. For some resolution M in dimension D we have $(M+1)^D$ points. Clearly for even moderate resolution this will not work.

In general, it is difficult to find the global optimal solution. Possible strategies are:

1. Greedy algorithm. Initialize the weights (say, all one), and iteratively improve the weights. In each iteration, adjust one weight while fixing all other weights.
2. Stochastic search algorithm. Use algorithms like simulated annealing and genetic algorithms.
3. Formulate the problem as a convex optimization problem. For example, one can search for optimal weights such that points in the same class are closer and points in different classes become far away.

Problem 3 *[20 points]*

Consider a binary logistic regression model where the training samples are *linearly separable*:

(a) Given n training examples $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, write down the negative log likelihood (as a loss function):

Solution: The loss function $\mathcal{L}$ is given by:

$$\mathcal{L}(\boldsymbol{w}) = -\sum_{i=1}^{n} \left[y_i \log(\sigma[\boldsymbol{w}^T \boldsymbol{x}]) + (1 - y_i) \log(1 - \sigma[\boldsymbol{w}^T \boldsymbol{x}]) \right]$$

Here, $\sigma[\cdot]$ denotes the sigmoid function, and x and w have been modified as in lecture such that the first column of x is entirely ones, and the first column of w is also known as b.

(b) Is this loss function convex? Provide your reasoning.

Solution: Yes, as the Hessian H is Positive Definite. Demonstration:

$$\begin{aligned}
\nabla \log \mathcal{L}(w) &= -\sum_{i=1}^{n} \left[y_i \frac{1}{\sigma[\boldsymbol{w}^T \boldsymbol{x}_i]} \frac{d}{d\boldsymbol{w}} \sigma[\boldsymbol{w}^T \boldsymbol{x}_i] + (1-y_i) \frac{1}{1-\sigma[\boldsymbol{w}^T \boldsymbol{x}_i]} \frac{d}{d\boldsymbol{w}} (1-\sigma[\boldsymbol{w}^T \boldsymbol{x}_i]) \right] \\
&= -\sum_{i=1}^{n} \left[y_i \frac{\sigma[\boldsymbol{w}^T \boldsymbol{x}_i](1-\sigma[\boldsymbol{w}^T \boldsymbol{x}_i])}{\sigma[\boldsymbol{w}^T \boldsymbol{x}_i]} \frac{d}{d\boldsymbol{w}} \boldsymbol{w}^T \boldsymbol{x}_i + (1-y_i) \frac{\sigma[\boldsymbol{w}^T \boldsymbol{x}_i](1-\sigma[\boldsymbol{w}^T \boldsymbol{x}_i])}{(1-\sigma[\boldsymbol{w}^T \boldsymbol{x}_i])} \frac{d}{d\boldsymbol{w}} \boldsymbol{w}^T \boldsymbol{x}_i \right] \\
&= -\sum_{i=1}^{n} \left[y_i (1-\sigma[\boldsymbol{w}^T \boldsymbol{x}_i]) \boldsymbol{x}_i + (1-y_i) \sigma[\boldsymbol{w}^T \boldsymbol{x}_i] \boldsymbol{x}_i \right] \\
&= \sum_{i=1}^{n} \left[(\sigma[\boldsymbol{w}^T \boldsymbol{x}_i] - y_i) \boldsymbol{x}_i \right]
\end{aligned}$$

$$\nabla^2 \log \mathcal{L}(w) = \nabla (\nabla \log \mathcal{L}(w))^T = \sum_{i=1}^{n} \sigma[\boldsymbol{w}^T \boldsymbol{x}_i](1-\sigma[\boldsymbol{w}^T \boldsymbol{x}_i]) \boldsymbol{x}_i \boldsymbol{x}_i^T = X^T \operatorname{diag}(\sigma[\boldsymbol{w}^T \boldsymbol{x}_i](1-\sigma[\boldsymbol{w}^T \boldsymbol{x}_i])) X$$

The diagonal of the matrix is strictly positive, making all of the eigenvectors strictly positive unless X is rank deficient in the feature dimension, meaning one feature is irrelevant (if this is the case there exists a linear subspace on which the data lies exactly, which has *lower* dimension than the full space; in the presence of continuous noise the probability of this goes to zero). Thus the Hessian is almost always Positive Definite, making this loss function convex in $\boldsymbol{w}$.

(c) Split the loss function by class:

$$\mathcal{L}(\boldsymbol{w}) = -\sum_{i=1}^{n} \left[y_i \log(\sigma[\boldsymbol{w}^T \boldsymbol{x}]) + (1-y_i) \log(1-\sigma[\boldsymbol{w}^T \boldsymbol{x}]) \right] = -\sum_{i=1, y_i=1}^{n} \left[\log(\sigma[\boldsymbol{w}^T \boldsymbol{x}]) \right] - \sum_{i=1, y_i=0}^{n} \left[\log(1-\sigma[\boldsymbol{w}^T \boldsymbol{x}]) \right]$$

If the data is linearly separable, then there exists a $\boldsymbol{w}$ such that the hyperplane it describes perfectly splits the data by class. Thus, for class $y_i = 1$, the sigmoid function will always be evaluated at a positive point (here w.l.o.g. we orient the hyperplane such that class $y_i = 1$ is "above" it). Furthermore, for the class $y_i = 0$, the sigmoid will always be evaluated at a negative point. However, since $1 - \sigma[\cdot] = \sigma[-\cdot]$, we can rearrange it such that the sigmoid is again being evaluated at a positive point.

$$-\sum_{i=1, y_i=1}^{n} \left[\log(\sigma[\boldsymbol{w}^T \boldsymbol{x}]) \right] - \sum_{i=1, y_i=0}^{n} \left[\log(1-\sigma[\boldsymbol{w}^T \boldsymbol{x}]) \right] = -\sum_{i=1, y_i=1}^{n} \left[\log(\sigma[\boldsymbol{w}^T \boldsymbol{x}]) \right] - \sum_{i=1, y_i=0}^{n} \left[\log(\sigma[-\boldsymbol{w}^T \boldsymbol{x}]) \right]$$

So we now need to minimize this function, which means we need to minimize $-\log(\sigma[\cdot])$, where the sigmoid is always evaluated at a positive point. So we want $\sigma[\cdot] \in (0,1)$ to approach 1, meaning that the magnitude of $\boldsymbol{w}^T x$ needs to be as large as possible. $\boldsymbol{x}$ is fixed, but this means that $\boldsymbol{w}$ will need to be very large, meaning that our algorithm will prefer that $\boldsymbol{w}$ "go to infinity".

Alternatively we can imagine that the sigmoid is optimal when it becomes 0–1 loss. This only occurs when $\boldsymbol{w}$ goes to infinity, as the 0–1 loss is discontinuous.

(d) So to not confuse ourselves with the previous steps, we do **not** refer the the regularized loss as $\mathcal{L}(\boldsymbol{w})$, only the fidelity term.

$$\frac{\partial}{\partial w_i} (\mathcal{L}(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_2^2) = -(\sigma[\boldsymbol{w}^T \boldsymbol{x}_i] - y_i) \boldsymbol{x}_i + \lambda \boldsymbol{w}_i$$

(e) By the linear property of partial derivative operators, the addition of a strictly convex function and a convex function is still strictly convex, so this function has one minimum; however, it does not ensure that this minimum is finite. If we evaluate this function at infinity, we find that it has infinite loss. Evaluating it at zero has finite loss, so clearly the minimum solution is finite (it must be less than or equal to the finite loss at zero).

Problem 4 *[15 points]*

Part (a) We should split on "Traffic" because it gives a perfect prediction of "Accident rate". The other cannot do perfect prediction.

Part (b) We can think about decision trees as partitioning the space of observations along each axis. If every feature is continuous and ordered we can transform T_1 into T_2 by taking each decision boundary, subtracting off the appropriate mean, and then dividing by the appropriate variance. Both trees have the same structure and same accuracy. In other words, linear transformation does not change informativeness of the features.

Part (c) Consider the difference between the Gini Index and Cross Entropy:

$$\begin{aligned} G - CE &= \sum_{k=1}^{K} [p_k(1 - p_k)] + \sum_{k=1}^{K} [p_k \log p_k] \\ &= \sum_{k=1}^{K} p_k(1 - p_k + \log p_k). \end{aligned}$$

Now examine the function $f(x) = 1 - x + \log(x)$, where the base of the log is less than or equal to e (the cross entropy is defined with base 2). Note that f is continuous on the positive real line.

Now consider the derivative $\frac{d}{dx} f = -1 + \frac{1}{x \log(a)}$ where a is the base of the log. This function is also continuous on the positive real line. For all $a \leq e$, $\log(a) \leq 1 \Rightarrow \frac{1}{x \log(a)} \leq 1$ for all $x \in (0, 1)$, and for $x = 1$, $\frac{1}{x \log(a)} = 1$. This implies that $\frac{d}{dx} f(x) > 0$ for $x \in (0, 1), a < e$ so f has no critical points in $(0, 1)$.

Note that $f(x) \to -\infty$ as $x \to 0+$ and consider $x = 1$. $f(x) = 0$, and has no previous critical points, so it cannot have any positive points (if f were to have a positive point, since it is continuous it must decrease to $f(0)$, but it then must have a negative derivative, meaning its derivative must have a zero, meaning it must have a critical point. Contradiction.).

Thus, $1 - p_k + \log p_k < 0$, meaning that $G - CE < 0$, meaning that the Gini Index is always less than the Cross Entropy.

Problem 5 *[35 points]*

Part 5) The solutions become more contiguous (the connected components of one class are larger, and have larger interiors), and the boundary is marginally smoother on a micro level, much smoother on a macroscopic level.