Problem 1 – Decision regions

1.1

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact</td>
<td>Appr</td>
</tr>
</tbody>
</table>
| ☒    | ☒       | 1-nearest neighbor classifier
| ☐    | ☐       | linear classifier
| ☐    | ☐       | quadratic classifier
| ☒    | ☐       | decision tree
| ☐    | ☒       | 2 layer neural network

Problem 2 – Stochastic gradient for additive model

2.1 \( L(\beta) = \frac{1}{N} \sum_{i=1}^{N} e^{-y^i \sum_{k=1:m} \beta_k b_k(x^i)} \)
2.2
\[
\frac{\partial L(y, f(x))}{\partial \beta_k} = \frac{\partial}{\partial \beta_k} e^{-y\sum_j \beta_j b_j(x)} = \prod_{j=1:m, j\neq k} e^{-y\beta_j b_j(x)} \frac{\partial e^{-y\beta_k b_k(x)}}{\partial \beta_k} = -y b_k(x) L(y, f(x))
\]
and \(\nabla_\beta L(y, f(x)) = -yb(x) L(y, f(x))\) where \(b(x)\) is the vector \([b_1(x) \ldots b_m(x)]^T\).

2.3 \(\frac{\partial \hat{L}(\beta)}{\partial \beta_j} = -\frac{1}{N} \sum_{i=1}^{N} e^{-y_i \sum_{k=1:m} \beta_k b_k(x^i)} y_i b_j(x^i)\)

2.4 \(\hat{L}\) is convex in \(\beta\) because it is a linear combination of the convex functions \(e^{\beta^T a^i}\) with \(a_i = [-y_i b_1(x^i) \ldots -y_i b_m(x^i)]^T\).

2.5 The gradient of \(\hat{L}\) is an average over the training set of \(\nabla_\beta L(y^i, f(x^i)) = d_i\). The expectation of \(d_i\) is \(\hat{L}\). Hence,

Initialize \(\beta = \bar{\beta} = 0\) (for example)

for \(t = 1, 2, \ldots T\)

1. pick \(i \in 1:n\) at random
2. \(d_i = \nabla_\beta L(y^i, f(x^i))\)
3. update \(\beta\) by

\[
\beta^{t+1} = \beta^t - \frac{c}{\lambda t} d_i
\]

\[
= \beta^t + \frac{c y^i e^\sum_{k=1}^{m} b_k b_k(x^i)}{\lambda t} \left[ \begin{array}{c} \cdots \\ b_k(x^i) \\ \ldots \end{array} \right]
\]  (2)

4. if \(t > (1 - \alpha)T\) \(\bar{\beta}^{t+1} \leftarrow \bar{\beta}^t + \beta^t\)

Return \(\bar{\beta}^{T+1}/(\alpha T)\)

**Problem 3 – Boosting**

3.1 Give the mathematical definition and name the following expressions

<table>
<thead>
<tr>
<th>Expression</th>
<th>Formula</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>(L_{01}(b))</td>
<td>(P_{XY}[Y \neq b(X)])</td>
<td>expected probability of error of (b), expected loss of (b)</td>
</tr>
<tr>
<td>(L_{01}(B))</td>
<td>(\min_{b \in B} L_{01}(b))</td>
<td>minimum training errors achievable by a classifier from (B)</td>
</tr>
<tr>
<td>(L_{01}(f))</td>
<td>(P_{XY}[Y \neq f(X)])</td>
<td>expected probability of error of (f), expected loss of boosted classifier (f)</td>
</tr>
</tbody>
</table>
3.2 We know that $\hat{L}_{01}(f^k) \leq \gamma^k$ with $\gamma = 2\sqrt{\delta(1-\delta)}$. To guarantee 0 error, it is sufficient that $\hat{L}_{01}(f^k) < 1/N$. Hence,

$$\gamma^k < \frac{1}{N} \Rightarrow k \ln \gamma < -\ln N \Rightarrow k > \frac{\ln N}{\ln \frac{1}{\gamma}} \Rightarrow k_0 = \left\lfloor \frac{\ln N}{\ln \frac{1}{\gamma}} \right\rfloor < \ln N$$

Problem 4 – Concave lower bound to convex function

By definition $F = \{(x, y), y \geq f(x)\}$ and $G = \{(x, y), y \leq g(x)\}$ are convex sets. Their interiors are disjoint, because $(x, y) \in F \cap G$ implies $y \leq g(x) \leq f(x) \leq y$ which in turn implies $y = f(x) = g(x)$, i.e. $(x, y) \in \partial F \cap \partial G$. (Moreover, the set $E \cap F$ must be a line segment, possibly unbounded or empty.) By the separating hyperplane theorem, there exists a hyperplane defined by $\{(x, y), H(x, y) = w^T x + w_0 y + b = 0\}$ so that

\begin{align*}
H(x, y) &\geq 0 \quad \text{on } F \quad (3) \\
H(x, y) &\leq 0 \quad \text{on } G \quad (4)
\end{align*}

W.l.o.g, $w_0 > 0$ (it is easy to see that $w_0$ cannot be 0), and since rescaling by a positive number does not affect (4) and (3), we can assume $w_0 = 1$. Therefore, on the hyperplane we have

$$w^T x + y + b = 0 \Rightarrow y = -w^T x - b = h(x). \quad (5)$$

The point $(x, f(x))$ is in $F$, and by (3) we have

$$0 \leq f(x) + w^T x + b = f(x) - h(x). \quad (6)$$

and similarly, $0 \geq g(x) - h(x)$, which completes the proof.

Problem 5 – Optimal discretization of a continuous variable

5.1 $X$ is completely determined by $Y$, therefore $H(P_{X|Y=y}) = 0$. $H(X|Y)$ is the expectation of $H(P_{X|Y=y})$ under $P_Y$, hence $H(X|Y) = 0$.

To write $P_{X|Y=y}$ explicitly: For any $y \in [0, 1)$ set $i(y)$ so that $y \in [a_{i-1}, a_i)$. There is a unique $i(y)$ for any $y$. Then,

$$P_{X|Y=y}(i) = \begin{cases} 1 & \text{if } i = i(y) \\ 0 & \text{otherwise} \end{cases}$$

5.2 We want to find $X$ to maximize $I(X, Y)$. But $I(X, Y) = H(X) - H(X|Y) = H(X)$. Thus, we need to maximize $H(X)$. Denote

$$p_i = \int_{a_{i-1}}^{a_i} f(y)dy = F(a_i) - F(a_{i-1}), \quad \text{for } i = 1 : m \quad (7)$$
The maximum of the entropy of an $m$-valued variable is attained when its distribution is uniform. Hence, $p_i = \frac{1}{m}$ and therefore $a_i$ are given by

$$a_i = F^{-1} \left( \frac{i}{m} \right).$$

(8)

A note: because we assumed that $f(y) > 0$, the inverse $F^{-1}$ is well defined. But (8) generalizes to any CDF for which $F^{-1}(i/m)$ is not the empty set, and in particular to any continuous CDF over any subset of $\mathbb{R}$. 