Elements of Machine Learning
Assignment 2

William LaCroix (7038732)
wila00001@stud.uni-saarland.de

Problem 2 (T, 10 Points). **Bayes-optimal classifier.**

The optimal misclassification error is achieved by the Bayes optimal classifier. This is the classifier that assigns every point X to its most likely class. That is, the Bayes optimal classifier predicts

$$\hat{y} = f^*(x) = \underset{y\in\{0,1\}}{argmaxP}(Y = y|X = x)$$

1. Consider a scalar feature $X \in \mathbb{R}^2$ and a binary random variable Y , for which.

$P(X|Y=0) = \begin{cases} \frac{1}{\pi r^2} & \|X\| \leq r \\ 0 & otherwise \end{cases}$,
$P(X|Y=1) = \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \sigma^2\begin{bmatrix}1&0\\0&1\end{bmatrix}\right) = \frac{1}{2\pi\sigma^2}exp\left(-\frac{\|X\|^2}{2\sigma^2}\right)$,
$P(Y=0) = cP(Y=1)$,
where $r, \sigma > 0$ and $0 < c < 1$ are parameters.

[6pts] Derive the Bayes optimal classifier for Y as a function of r and σ.

Bayes' Formula: $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$

Then: $\frac{P(X|Y=1)P(Y=1)}{P(X)} \geq \frac{P(X|Y=0)P(Y=0)}{P(X)}$ becomes $P(X|Y=1)P(Y=1) \geq P(X|Y=0)P(Y=0) \equiv$

$P(X|Y=1)P(Y=1) \geq P(X|Y=0)cP(Y=1) \equiv P(X|Y=1) \geq P(X|Y=0) * c \equiv$

$\frac{1}{2\pi\sigma^2}exp\left(-\frac{\|X\|^2}{2\sigma^2}\right) \geq \frac{1}{\pi r^2} * c \equiv exp\left(-\frac{\|X\|^2}{2\sigma^2}\right) \geq \frac{2c\sigma^2}{r^2} \equiv -\frac{\|X\|^2}{2\sigma^2} \geq log\left(\frac{2c\sigma^2}{r^2}\right) \equiv \|X\|^2 \geq -2\sigma^2 log\left(\frac{2c\sigma^2}{r^2}\right)$

Thus: $\hat{y} = f^*(x) = \underset{y\in\{0,1\}}{argmaxP}(Y = y|X = x) = \begin{Bmatrix} 1 & P(Y=1|X) \geq P(Y=0|X) \\ 0 & otherwise \end{Bmatrix} = \begin{Bmatrix} 1 & \|X\|^2 \geq -2\sigma^2 log\left(\frac{2c\sigma^2}{r^2}\right) \\ 0 & otherwise \end{Bmatrix}$

[2pts] Draw the decision boundary for $\sigma = 1, r = e\sqrt{2} \approx 3.84$ and $c = exp(-\frac{1}{3})$; explain your observations. What will happen to the decision boundary, if we increase c while keeping all the other parameters fixed?

Evaluating $\|X\|^2 \geq -2\sigma^2 log\left(\frac{2c\sigma^2}{r^2}\right)$ with $\sigma = 1, r = e\sqrt{2} \approx 3.84$ and $c = exp(-\frac{1}{3})$, we get
$Y = \begin{cases} 1 & \|X\|^2 \geq 4.66226177194 \\ 0 & otherwise \end{cases}$

As c approaches 1, the decision boundary lowers from $\|X\|^2 \geq 4.66226177194$ to a minimum of $\|X\|^2 \geq 2$, since $\lim_{c\to 1} -2\sigma^2 log\left(\frac{2c\sigma^2}{r^2}\right) = 2$. At the same time, $P(Y=0)$ approaches $P(Y=1)$

2. [2pts] Given that the Bayes optimal classifier has the lowest misclassification error among all classifiers, why do we need any other classification method?

The Bayes optimal classifier requires that we estimate the class prior probabilities based on the data, which will have varying distributions (Gaussian or otherwise), and may or may not have the same variance. Our different classification methods make different assumptions about the class prior probability distributions, and often make bias-variance trade-offs in order to best approximate the probabilities, for any given data set.