DATA 1010: Homework 6

This is a "take home" exam. If you need help then you should talk to me, Sam Watson, or one of the TAs, and not to each other. We will be generous with our answers.

You must write up your solutions on your own. There will definitely be related questions on Friday's exam, so it would be wise to make every effort to take these problems seriously and to use them as a study opportunity. This assignment will make up 40% of your grade on the classification material.

Due by 4:45 PM on *Wednesday, December 13*. Put your solutions into the DATA 1010 box in the Science Library.

Solutions will be posted within about an hour after the deadline.

Curse of Dimensionality. Consider a two-category classification problem, based on features $\vec{X} \in \mathbb{R}^d$ and classifications $Y \in \{1, 2\}$. The prior class probabilities are $\pi_1 = 0.4$ and $\pi_2 = 0.6$. Assume that the d features are independent given the category:

$$\begin{aligned} f_1(\vec{x}) &= \prod_{i=1}^{d} f_{1,i}(x_i) \\ &= \prod_{i=1}^{d} g(x_i; \mu_{1,i}, 1) \end{aligned}$$

$$\begin{aligned} f_2(\vec{x}) &= \prod_{i=1}^{d} f_{2,i}(x_i) \\ &= \prod_{i=1}^{d} g(x_i; \mu_{2,i}, 1) \end{aligned}$$

where $g(x; \mu, \sigma^2)$ is the density of the normal distribution with mean μ and variance σ^2:

$$g(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The model specification is completed by assigning values to the class-conditional means. For each category $c = 1, 2$ and feature $i = 1, 2, \ldots, d$:

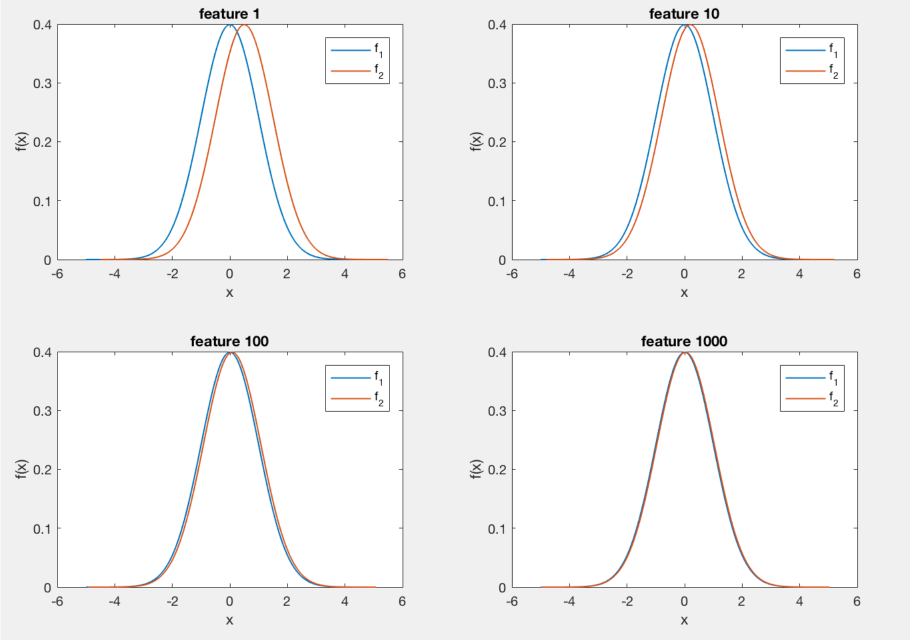
$$\mu_{1,i} = 0 \quad \text{and} \quad \mu_{2,i} = \frac{0.5}{i^{0.4}}$$

As usual, the goal is to use a training set consisting of feature vectors and their classifications $\{\vec{x}(k), y(k)\}_{k=1}^{n}$ to train a classifier $h : \mathbb{R}^d \to \{1, 2\}$, $h = h(x; \{\vec{x}(k), y(k)\}_{k=1}^{n})$.

Pretend like we don't know d, we don't know the means, and we don't make the naive Bayes assumption that the features are independent given the category (even though they are). We have lots of features lying around that we could observe (in fact, an infinite number, one for every integer, $i = 1, 2, \ldots$), but choose to use what we deem to be the d most relevant, say $x_1, x_2, \ldots, x_d$. Since we know nothing about $f_1(x_1, \ldots, x_d)$ or $f_2(x_1, \ldots, x_d)$ we will take a non-parametric approach and build a k nearest-neighbor classifier.

Notice that every feature carries information that could be useful in deciding the classification. Although all of the features have standard deviation one, independent of the category, the mean of every feature is different from one category to the other. In fact, in category one all of the features have zero means, and in category two they all have positive means. But the difference between the two diminishes as the index of the feature increases: $\mu_{2,i} - \mu_{1,i} = 0.5/i^{0.4} \to 0$ as $i \to \infty$.

To get a feeling for the information available from the features, the figure below plots the two densities, $f_{1,i}$ and $f_{2,i}$ together, for each of a selection of features i.



Evidently, there is not a lot of information in any one feature, but perhaps when d is large there is enough collective information to get good performance. How are the feature observations best combined to build a classifier? If we knew the distributions, then the very best way would be to use the Neyman-Pearson lemma and threshold on the likelihood ratio:

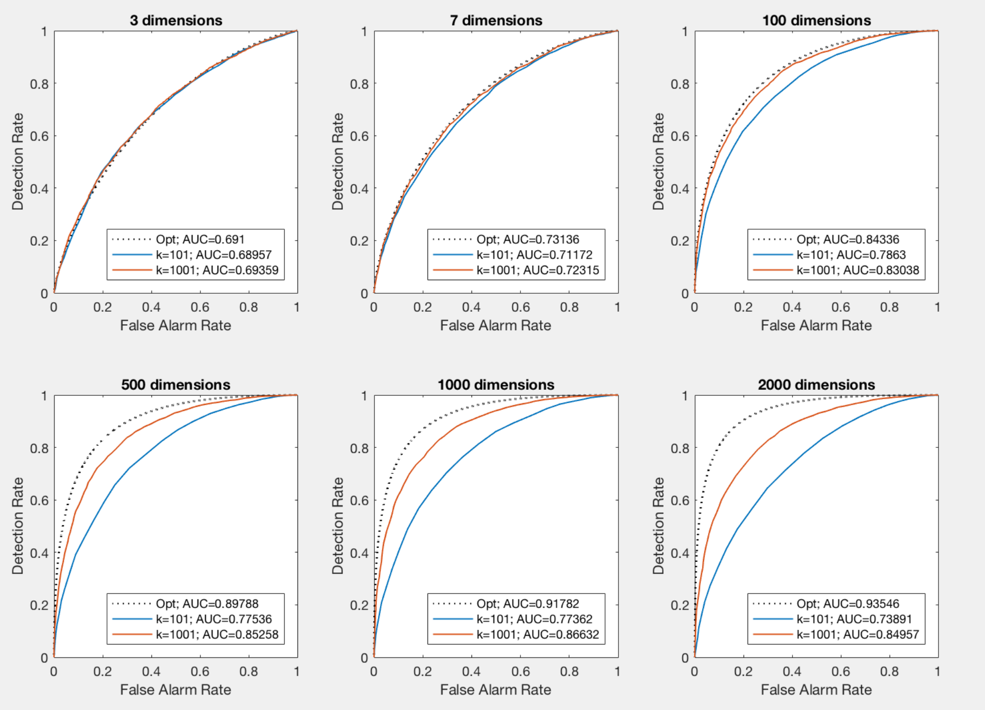
$$h_t(\vec{x}) = h_t(x_1, \dots, x_d) \triangleq \begin{cases} 1 & \text{if } \frac{f_2(\vec{x})}{f_1(\vec{x})} < t \\ \\ 2 & \text{if } \frac{f_2(\vec{x})}{f_1(\vec{x})} \geq t \end{cases} \tag{1}$$

We will explore two questions about performance, one theoretical and the other experimental:

(i) How well would the Neyman-Pearson classifier perform at high values of d?

(ii) For a given sample size, how well does the k nearest neighbor classifier (kNN) perform as more features are added (d gets larger)?

Let's start by looking at some analytic and some experimental results:



There are three ROC curves in each subplot: One for the theoretically optimal classifier (likelihood ratio test—see (1) above), and one for each of two values of k, $k = 101$ and $k = 1001$. The latter two were generated by experimenting with data generated from the model. In each experiment, 3,000 samples (i.e. pairs of $\vec{x}$ and the true category y) were used for training, and 10,000 additional samples were used for (i) testing, (ii) generating the ROC curves for $k = 101$ and $k = 1001$, and (iii) computing the areas under these curves ("AUC").

Observe that

(a) The area under the optimal ROC curve increases with the addition of more features. This can't be a surprise, since every feature adds at least some information to the decision.

(b) For both k =101 and k =1,001, performance begins by increasing with the addition of more features, but, eventually, further additions make the performance worse.

(c) Performance in low dimensions is indistinguishable from optimal; performance in high dimensions is very far from optimal.

1. How far towards perfect (AUC=1) would the optimal classifier get if we were to continue adding features? It appears to be almost unchanged by going from 500 to 1,000 features, so it would be reasonable to guess that there is little or nothing further to be gained. You might be surprised, then, to learn that for this problem the Neyman-Pearson (likelihood ratio) classifier is asymptotically perfect. Prove it!

   Hint: follow these steps...

   (i) The ROC curve is the set of points $\{(FAR(t), DR(t)) : t \in (0, \infty)\}$. Use equation (1) and the definitions of $DR(t)$ and $FAR(t)$ to show that

$$DR(t) = \mathbb{P}\left(\frac{f_2(\vec{X})}{f_1(\vec{X})} \geq t | \vec{X} \sim f_2\right)$$

$$FAR(t) = \mathbb{P}\left(\frac{f_2(\vec{X})}{f_1(\vec{X})} \geq t | \vec{X} \sim f_1\right)$$

It seems that the likelihood ratio, $L \triangleq \frac{f_2(\vec{X})}{f_1(\vec{X})}$, is the key term. We will look for a simple expression for $\log(L)$.

(ii) Since $\mu_{1,i} = 0$ for all i, there will be no confusion if we simply write μ_i for $\mu_{2,i}$. Later, we will substitute the actual values, $\mu_i = \frac{0.5}{i^{0.4}}$, for $\mu_i,\ i = 1, 2, \ldots$, but for now it will be more convenient to stick with symbols. With this convention, we can write

$$\frac{f_2(\vec{X})}{f_1(\vec{X})} = \prod_{i=1}^{d} \frac{g(X_i; \mu_i, 1)}{g(X_i; 0, 1)}$$

Show that for each i there exists α_i and β_i such that

$$\log\left(\frac{g(X_i; \mu_i, 1)}{g(X_i; 0, 1)}\right) = \alpha_i X_i + \beta_i$$

Write down explicit expressions for α_i and β_i as functions of μ_i.

(iii) Let W be the random variable

$$W \triangleq \log\left(\frac{f_2(\vec{X})}{f_1(\vec{X})}\right)$$

Use the previous calculation to show that $W = \sum_{i=1}^{d} \alpha_i X_i + \sum_{i=1}^{d} \beta_i$. From this, and the independence of the features $X_1, X_2, \ldots$, argue that

$$W \sim \mathcal{N}(\mu_d^{(1)}, \sigma_d^2) \quad \text{when } \vec{X} \sim f_1$$

$$W \sim \mathcal{N}(\mu_d^{(2)}, \sigma_d^2) \quad \text{when } \vec{X} \sim f_2$$

where $\mu_d^{(1)}$, $\mu_d^{(2)}$, and σ_d^2 are functions of the means $\mu_1, \mu_2, \ldots, \mu_d$. Derive explicit expressions for all three.

(iv) For each of the two cases, $\vec{X} \sim f_1$ and $\vec{X} \sim f_2$, rewrite W in terms of a standard normal random variable $Z \sim \mathcal{N}(0, 1)$:

$$W = a_d^{(1)} Z + b_d^{(1)} \quad \text{when } \vec{X} \sim f_1$$

$$W = a_d^{(2)} Z + b_d^{(2)} \quad \text{when } \vec{X} \sim f_2$$

Derive explicit expressions for $a_d^{(1)}$, $b_d^{(1)}$, $a_d^{(2)}$, and $b_d^{(2)}$ in terms of $\mu_1, \mu_2, \ldots, \mu_d$.

(v) Getting back to DR and FAR, show that

$$DR(t) = \mathbb{P}\left(Z \geq \frac{\log(t) - b_d^{(2)}}{a_d^{(2)}}\right)$$

$$FAR(t) = \mathbb{P}\left(Z \geq \frac{\log(t) - b_d^{(1)}}{a_d^{(1)}}\right)$$

(vi) Finally, for every $i = 1, \ldots, d$ substitute $\frac{0.5}{i^{0.4}}$ for μ_i, and then show that for every $t \in (0, \infty)$

$$\lim_{d\to\infty} DR(t) = 1 \quad \text{and} \quad \lim_{d\to\infty} FAR(t) = 0$$

For the last step, this will be helpful: if

$$\gamma_d = \sum_{i=1}^{d} \frac{1}{i^{0.8}}$$

then $\lim_{d\to\infty} \gamma_d = \infty$.[1]

2. Classification is often based on some function of the feature vector, $G(\vec{x})$, which gives evidence one way or another for the correct classification: large values of G suggest category 2 over category 1, and vice versa. The likelihood ratio is an important example ($G(\vec{x}) = f_2(\vec{x})/f_1(\vec{x})$), but many (perhaps even most) classifiers have this relationship to some such a function. In these cases the classification function, h, is usually defined in terms of a threshold on G:

$$h(\vec{x}) = h_t(\vec{x}) = \begin{cases} 1 & \text{if } G(\vec{x}) < t \\ \\ 2 & \text{if } G(\vec{x}) \geq t \end{cases} \tag{2}$$

If G is continuous, then there is an elegant and disarmingly simple interpretation of the area (AUC) under the associated ROC curve: If $\vec{V} \sim f_1$ and $\vec{W} \sim f_2$ are chosen independently, then

$$\text{AUC} = \mathbb{P}(G(W) \geq G(V))$$

Samples from category 2 should typically produce larger values of G than those produced by samples from category 1. The more likely this is to occur, the greater the area under the curve.[2]

According to part (iii) of the previous problem, if we choose V and W independently using

$$V \sim \mathcal{N}(\mu_d^{(1)}, \sigma_d^2)$$

$$W \sim \mathcal{N}(\mu_d^{(2)}, \sigma_d^2)$$

[1] More generally, $\sum_{i=1}^{d} \frac{1}{i^q} \to \infty$ for all $q \in [0, 1]$, but it has a finite limit for all $q > 1$.

[2] Think of it this way: G is a statistic on which we are basing an hypothesis test. $Y = 1$ is the null hypothesis and $Y = 2$ is the alternative. The p-value of the test is the false alarm rate, which is determined by the size of the critical region, $G(\vec{X}) \geq t$. The larger the value of t the smaller the p-value.

then the AUC for the likelihood ratio test is just $\mathbb{P}(W \geq V)$. Let $U = W - V$. Then U is a Gaussian random variable and AUC $= \mathbb{P}(U \geq 0)$. Evidently, $\mathbb{E}[U] = \mu_d^{(2)} - \mu_d^{(1)}$ and $\mathbb{V}[U] = 2\sigma_d^2$.

Using these relationships, show that as $d \to \infty$ AUC$\to 1$. (Hint: you might find it easier to rewrite the event $U \geq 0$ in terms of a standard normal random variable, Z.)