ELE 535: Machine Learning and Pattern Recognition
Homework 3

Zachary Hervieux-Moore

Monday 8th October, 2018

Exercise 1: In Chapter 1 we developed the binary MAP classifier

$$f(x) = \begin{cases} 1, & \text{if } \ln\left(\frac{p_1(x)}{p_0(x)}\right) > \ln\left(\frac{p(0)}{p(1)}\right) \\ 0, & \text{otherwise} \end{cases}$$

This is based on an underlying generative model in which $p(k)$ is the prior probability of class k, and $p_k(x) = p(x|k)$ the conditional density of the datum x given that the class k, $k \in \{0, 1\}$. Now assume that $x \in \mathbb{R}^n$, and that $p_k(x)$ is a multivariate Gaussian density

$$p_k(x) = \frac{1}{(2\pi)^{n/2}} \frac{1}{|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}$$

a) Determine the resulting MAP classifier in its simplest form.

b) Determine the form of the decision boundary for this classifier.

c) An empirical version of the MAP classifier is obtained by using training data to estimate any unknown parameters. Then using these estimates in the MAP classifier. Compare this empirical Bayes classifier to the nearest centroid classifier.

Answer:

a) Substituting the conditional probabilities into the MAP classifier given, we get

$$\ln\left(\frac{\frac{1}{(2\pi)^{n/2}} \frac{1}{|\sigma^2 I_n|^{1/2}} e^{-\frac{1}{2\sigma^2}(x-\mu_1)^T(x-\mu_1)}}{\frac{1}{(2\pi)^{n/2}} \frac{1}{|\sigma^2 I_n|^{1/2}} e^{-\frac{1}{2\sigma^2}(x-\mu_0)^T(x-\mu_0)}}\right) > \ln\left(\frac{p(0)}{p(1)}\right)$$

$$\|x-\mu_0\|_2^2 - \|x-\mu_1\|_2^2 > 2\sigma^2 \ln\left(\frac{p(0)}{p(1)}\right)$$

Which I have left in the form simplest to me. This gives the classifier

$$f(x) = \begin{cases} 1, & \text{if } \|x-\mu_0\|_2^2 - \|x-\mu_1\|_2^2 > 2\sigma^2 \ln\left(\frac{p(0)}{p(1)}\right) \\ 0, & \text{otherwise} \end{cases}$$

b) The boundary occurs when

$$\|x - \mu_0\|_2^2 - \|x - \mu_1\|_2^2 = 2\sigma^2 \ln\left(\frac{p(0)}{p(1)}\right)$$

Expanding the terms and simplifying the LHS yields

$$x^T(\mu_1 - \mu_0) = \sigma^2 \ln\left(\frac{p(0)}{p(1)}\right) + \frac{1}{2}\|\mu_1\|_2^2 - \frac{1}{2}\|\mu_0\|_2^2$$

Which of course is a linear boundary that has an affine shift.

c) The empirical version of the classifier is

$$f(x) = \begin{cases} 1, & \text{if } \|x - \hat{\mu}_0\|_2^2 - \|x - \hat{\mu}_1\|_2^2 > 2\sigma^2 \ln\left(\frac{p(0)}{p(1)}\right) \\ 0, & \text{otherwise} \end{cases}$$

This corresponds precisely to the nearest centroid classifier when $p(0) = p(1)$. That is, our priors can bias which centroid we put a larger emphasis on.

Exercise 2: (Naive Bayes classifier) Derive the MAP classifier when the conditional probability density $p_k(x)$ is the multivariate Gaussian density as in 1) with Σ_k a diagonal matrix, $k = 0, 1$. As before, the prior probability $p(k)$ of class $k \in \{0, 1\}$ is given

a) Determine the resulting MAP classifier in its simplest form.

b) Determine the form of the decision boundary for this classifier.

c) An empirical version of the MAP classifier is obtained by using training data to estimate any unknown parameters. Then using these estimates in the MAP classifier. This yields the Naive Bayes classifier.

Answer: We duplicate the steps from the question 1,

a) Substituting the conditional probabilities into the MAP classifier given, we get

$$\begin{aligned}
&\ln\left(\frac{\frac{1}{(2\pi)^{n/2}}\frac{1}{|\Sigma_1|^{1/2}}e^{-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)}}{\frac{1}{(2\pi)^{n/2}}\frac{1}{|\Sigma_0|^{1/2}}e^{-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)}}\right) > \ln\left(\frac{p(0)}{p(1)}\right)\\
&\implies (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) - (x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\\
&\qquad > 2\ln\left(\frac{p(0)}{p(1)}\right) + 2\ln|\Sigma_1|^{1/2} - 2\ln|\Sigma_0|^{1/2}
\end{aligned}$$

This gives the classifier

$$f(x) = \begin{cases} 1, & \text{if } (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) - (x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) \\ & \qquad > 2\ln\left(\frac{p(0)}{p(1)}\right) + 2\ln|\Sigma_1|^{1/2} - 2\ln|\Sigma_0|^{1/2} \\ 0, & \text{otherwise} \end{cases}$$

b) The boundary occurs when

$$\begin{aligned}
&(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) - (x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\\
&\qquad = 2\ln\left(\frac{p(0)}{p(1)}\right) + 2\ln|\Sigma_1|^{1/2} - 2\ln|\Sigma_0|^{1/2}
\end{aligned}$$

Expanding the terms and simplifying the LHS yields

$$\begin{aligned}
&x^T \left(\Sigma_0^{-1} - \Sigma_1^{-1}\right) x + 2x^T(\Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0) \\
&+\mu_0^T\Sigma_0^{-1}\mu_0 - \mu_1^T\Sigma_1^{-1}\mu_1 = \sigma^2 \ln\left(\frac{p(0)}{p(1)}\right) + \frac{1}{2}\|\mu_1\|_2^2 - \frac{1}{2}\|\mu_0\|_2^2
\end{aligned}$$

Which of course is a qudratic boundary.

c) The empirical version of the classifier is

$$f(x) = \begin{cases} 1, & \text{if } (x - \hat{\mu}_0)^T\hat{\Sigma}_0^{-1}(x - \hat{\mu}_0) - (x - \hat{\mu}_1)^T\hat{\Sigma}_1^{-1}(x - \hat{\mu}_1) \\ & \qquad > 2\ln\left(\frac{p(0)}{p(1)}\right) + 2\ln|\hat{\Sigma}_1|^{1/2} - 2\ln|\hat{\Sigma}_0|^{1/2} \\ 0, & \text{otherwise} \end{cases}$$

This corresponds precisely to the Naive Bayes estimator as reshuffling the above will yield $p_1(x)p(1) > p_0(x)p(0)$ where the conditional probabilities use the empirical parameters.

Exercise 3: Let $X \in \mathbb{R}^{n \times m}$ be a data matrix with data items stored in the columns of X. Show that the set of nonzero eigenvalues of XX^T is the same as the set of nonzero eigenvalues of X^TX.

Answer: Let the SVD of X be $U\Sigma V^T$ then the SVD of X^T is $V\Sigma U^T$. Then we have

$$\begin{aligned} XX^T &= U\Sigma U^T \\ X^TX &= V\Sigma V^T \end{aligned}$$

Then the eigenvectors of the above are precisely the columns vectors of U and V respectively. Using the i^{th} column yields

$$\begin{aligned} XX^Tu_i &= U\Sigma U^Tu_i = U\Sigma e_i = U\sigma_i e_i = \sigma_i u_i \\ X^TXv_i &= V\Sigma V^Tv_i = V\Sigma e_i = V\sigma_i e_i = \sigma_i v_i \end{aligned}$$

Since this holds for all i amd U and V have the same number of columns, the result follows.

Exercise 4: Some additional properties of singular values. Let $\sigma_i(A)$ denote the i^{th} singular value of the matrix A, $i \in [1:r]$ with $r = \text{rank}(A)$. Prove the following.

a) If λ is an eigenvalue of $A \in \mathbb{R}^{n\times n}$, then $|\lambda| \leq \sigma_1(A)$.

b) For $A \in \mathbb{R}^{m\times n}$, $|A_{i,j}| \leq \sigma_1(A)$, $i \in [1:m]$, $j \in [1:n]$.

c) For $\alpha \in \mathbb{R}$ and $A \in \mathbb{R}^{m\times n}$, $\sigma_i(\alpha A) = |\alpha|\sigma_i(A)$, $i \in [1:r]$.

Answer:

a) If λ is an eigenvalue, we have $Av = \lambda v$ for some eigenvector v with norm equal to 1. We also have $\sqrt{v^T A^T A v} = |\lambda|$. Then, by definition

$$\sigma_1(A) = \max_{\|x\|_2=1} \|Ax\|_2 = \max_{\|x\|_2=1} \sqrt{x^T A^T A x} \geq \sqrt{v^T A^T A v} = |\lambda|$$

b) Doing similar steps as the previous part, we lower bound the maximum by using a canonical vector e_i

$$\sigma_1(A) = \max_{\|x\|_2=1} \|Ax\|_2 \geq \sqrt{e_i^T A^T A e_i} = \sqrt{A_{i,:}^T A_{:,i}} = \sqrt{\sum_{j=1}^{n} A_{i,j}^2} \geq |A_{i,j}|$$

Since this holds for all canonical vectors e_i, the result follows.

c) Let A have the SVD $A = U\Sigma V^T$ then we have $\alpha A = U(\alpha\Sigma)V^T$. This implies

$$(\alpha A)^T(\alpha A) = U(\alpha^2\Sigma)U^T = \alpha^2 U\Sigma U^T$$

and so we have that $\sigma_i(\alpha A)^2 = \alpha^2\sigma_i(A)^2$. Taking the square root gives the result $\sigma_i(\alpha A) = |\alpha|\sigma_i(A)$.

Exercise 5: Nuclear Norm. Let $A \in \mathbb{R}^{m\times n}$ have rank r with $r \leq q = \min(m,n)$. Define the nuclear norm of A by $\|A\|_* = \sum_{i=1}^r \sigma_i(A)$.

a) Find $B \in \mathbb{R}^{m\times n}$ that maximizes $\langle A, B\rangle$ subject to $\sigma_1(B) \leq 1$.

b) Show that $\|A\|_* = \max_{\|C\|_2 \leq 1} \langle A, C\rangle$.

c) Show that $\|\cdot\|_*$ is a norm on $\mathbb{R}^{m\times n}$.

Answer:

a) First, let me show that we can assume that A is diagonal. Let A have SVD $A = U\Sigma V^T$. Then

$$\langle A, B\rangle = \langle U\Sigma V^T, B\rangle = \langle \Sigma, U^T BV\rangle = \langle \Sigma, C\rangle$$

Where we can change B to C without changing the constraint of $\sigma(B) \leq 1$ because U and V are unitary and the spectral norm is invariant under unitary transformations. This simplifies the problems to

$$\max_{\sigma(C)\leq 1} \langle \Sigma, C\rangle$$

for a diagonal Σ. But having a diagonal Σ simplifies the trace to

$$\langle \Sigma, C\rangle = \sum_i \sum_j \Sigma_{i,j} C_{i,j} = \sum_i \Sigma_{i,i} C_{i,i}$$

Using question 4a) which says $|C_{i,i}| \leq \sigma_1(C) = 1$, we maximize the objective by picking $C_{i,i} = 1$. Or pick $C = I_r$. Transforming this back to B gives $B = UV^T$.

b) This is trivial by continuing with the summation at the end of the previous part and our choice of C. This yields

$$\max_{\|B\|_2\leq 1} \langle A, B\rangle = \max_{\sigma(B)\leq 1} \langle A, BC\rangle = \max_{\sigma(C)\leq 1} \langle \Sigma, C\rangle = \sum_i \Sigma_{i,i} = \sum_{i=1}^r \sigma_i(A)$$

c) The norm follows easily from the fact that the trace is an inner product and the maximum preserves all the inner product properties. However, it is presented here for completeness. The three properties of the norms are satisfied:

i) Triangle inequality,

$$\begin{aligned}\|A+B\|_* &= \max_{\|C\|_2 \leq 1} \langle A+B, C\rangle = \max_{\|C\|_2 \leq 1} \langle A, C\rangle + \langle B, C\rangle \\ &\leq \max_{\|C\|_2 \leq 1} \langle A, C\rangle + \max_{\|D\|_2 \leq 1} \langle B, D\rangle = \|A\|_* + \|B\|_*\end{aligned}$$

ii) Homogeneity, using question 4c)

$$\|\alpha A\|_* = \sum_{i=1}^{r} \sigma_i(\alpha A) = |\alpha| \sum_{i=1}^{r} \sigma_i(A) = |\alpha| \|A\|_*$$

iii) Positive definiteness,

$$\begin{aligned}\|A\|_* = 0 &\implies \sum_{i=1}^{r} \sigma_i(A) = 0 \\ &\implies \sigma_i(A) = 0 \quad \forall i \\ &\implies A = 0\end{aligned}$$