ML HW #3

Joshua Shapiro
11/4/16

1)1) Definition 1: $K(x,x')$ is a kernel if it can be written as an inner product $\phi(x)^T\phi(x')$ for some feature mapping $x \to \phi(x)$

Definition 2: $K(x,x')$ is a kernel if for any finite set of training examples $x_1,\ldots,x_n$, the $n\times n$ matrix K such that $K_{ij} = K(x_i,x_j)$ is positive semidefinite.

Show definition 1 implies 2.

Positive semidefinite implies the dot product of $M\alpha$ & α (where α is a vector) is ≥ 0.

$G_{ij} = k(x_i,x_j) = \langle \phi(x_i), \phi(x_j)\rangle$ for $i,j = 1,\ldots n$

$\alpha = \alpha_1 \ldots \alpha_n$

$$\alpha^T G\alpha = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j G_{ij} = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \langle \phi(x_i), \phi(x_j)\rangle = \left\langle \sum_{i=1}^n \alpha_i \phi(x_i), \sum_{j=1}^n \alpha_j \phi(x_j)\right\rangle = \left\| \sum_{i=1}^n \alpha_i \phi(x_i)\right\|^2 \geq 0$$

Therefore, G_{ij} is positive semidefinite, proving definition 2. □

1)2)a) Let $\phi^1(x)$ & $\phi^2(x)$ be the feature vectors corresponding to kernels $K_1(x,x')$ and $K_2(x,x')$. These feature vectors may be of different length. Show that the product is a kernel.

Let a be the feature vector of K_1, b of K_2

$\Rightarrow K_1(x,x') = a(x)^T a(x')$ & $K_2(x,x') = b(x)^T b(x')$

$a = \left[\;\right]\}m \quad b = \left[\;\right]\}n$

a has length m, b has length n

$$K_3 = K_1(x,x')K_2(x,x') = \left(\sum_{i=1}^m a_i(x)a_i(x')\right)\left(\sum_{i=1}^n b_i(x)b_i(x')\right) = \sum_{i=1}^m \sum_{j=1}^n \left(a_i(x)b_j(x)a_i(x')b_j(x')\right)$$

$$= \sum_{i=1}^m \sum_{j=1}^n c_{ij}(x)c_{ij}(x') = c(x)^T c(x')$$ where c is an m by n vector, so $c_{mn}(z) = a_m(z)b_n(z)$

$\Rightarrow K_3(x,x') = c(x)^T c(x')$ □

1)2)b) Build the following kernel: $K(x,x') = \left(1 + \left(\frac{x}{\|x\|}\right)^T\left(\frac{x'}{\|x'\|}\right)\right)^3$. Assume $K_0 = 1$ & $K_1 = x^T x'$

1) Scale: $f(x)K_1(x,x')f(x')$ where $f(x) = 1/\|x\|$

$$\frac{1}{\|x\|}x^T x'\frac{1}{\|x'\|} = \frac{x}{\|x\|}^T\frac{x'}{\|x'\|}$$

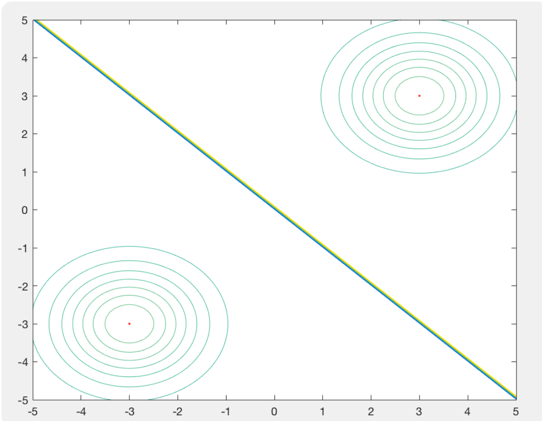
2) Add K_0

$$\Rightarrow 1 + \frac{x}{\|x\|}^T\frac{x'}{\|x'\|}$$

3) Multiply $K \cdot K \cdot K$

$$\Rightarrow \left(1 + \frac{x}{\|x\|}^T\frac{x'}{\|x'\|}\right)^3$$ □

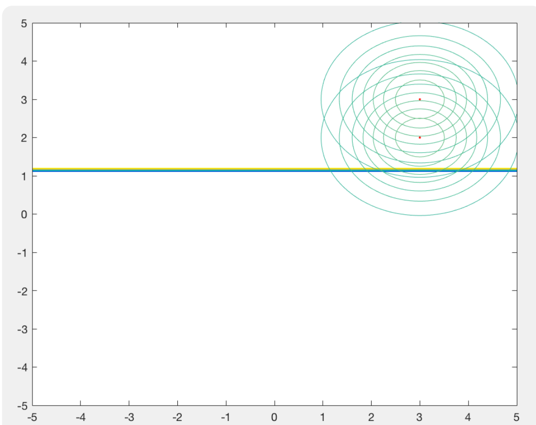
2)a)i) Linear decision boundary between the means of the two gaussians



Name ▲	Value
mixture	*1x1 struct*
mu1	[3,3]
mu2	[-3,-3]
sigma1	[1,0;0,1]
sigma2	[1,0;0,1]
wts	[0.5000,0.5000]

Gaussians with the same shape and weight that are apart will have a linear boundary.

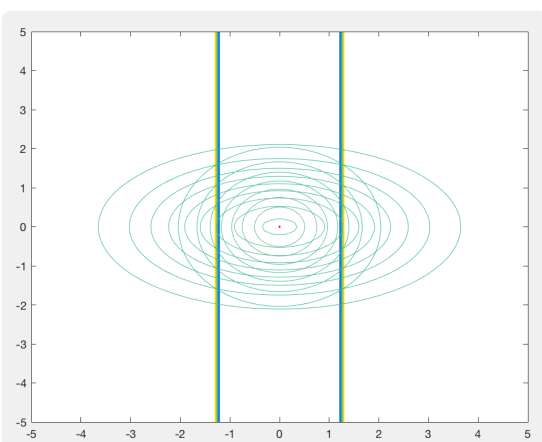
ii) Linear decision boundary where both means are on the same side of the decision boundary



Name ▲	Value
mixture	*1x1 struct*
mu1	[3,3]
mu2	[3,2]
sigma1	[1,0;0,1]
sigma2	[1,0;0,1]
wts	[0.8000,0.2000]

If two are close together, and one has a substantially higher weight, both means will be on the same side of the line.

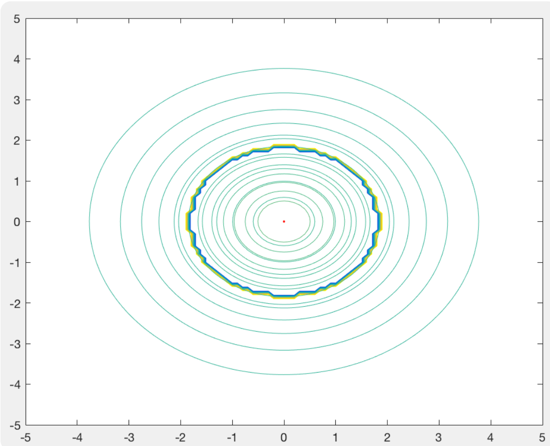
iii) A non-continuous decision boundary (one of the classes is represented by 2 disconnected regions)



Name ▲	Value
mixture	*1x1 struct*
mu1	[0,0]
mu2	[0,0]
sigma1	[3,0;0,1]
sigma2	[1,0;0,1]
wts	[0.5000,0.5000]

If the sigma matrix of one gaussian is modified to stretch the gaussian in the X direction and the other gaussian stays the same and sheres the mean, the stretched area will cause 2 decision boundaries to form.

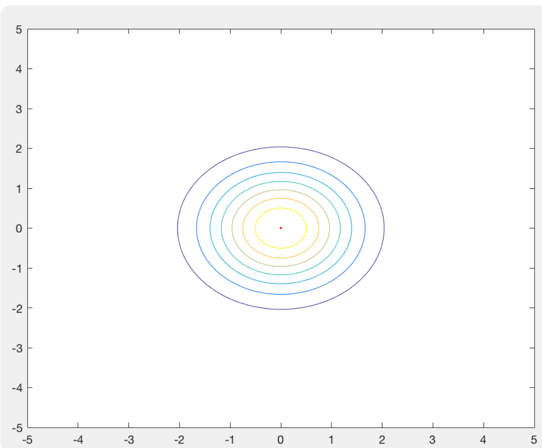
iv) A circular decision boundary



Name ▲	Value
mixture	*1x1 struct*
mu1	[0,0]
mu2	[0,0]
sigma1	[3,0;0,3]
sigma2	[1,0;0,1]
wts	[0.5000,0.5000]

Both gaussians are on top of each other, but one has larger sigma values. The outside becomes one decision, the inside becomes another.

v) No decision boundary - the entire plane is on one decision region



Name ▲	Value
mixture	*1x1 struct*
mu1	[0,0]
mu2	[0,0]
sig 1x2 double	[1,0;0,1]
sigma2	[1,0;0,1]
wts	[0.5000,0.5000]

If both gaussians have the same parameters, there can be no decision boundary since every point has the same probability of being in gaussian 1 as it does in gaussian 2

2)b) Show when k=2 the softmax model reduces to logistic regression.

Softmax probabilities: $Pr(y=i|x) = \dfrac{\exp(-z_i)}{\sum_{j=1}^{k} \exp(-z_j)}$ $\quad z_i = w_{i0} + \sum_j w_{ij} x_j = w_{i0} + w_i^T x$

Logistic Regression: $Pr(y=i|x) = \dfrac{1}{1 + e^{-y(w^T x + w_0)}}$ $\quad w = w_1, \dots, w_d$

For 2 class softmax and d features, there are 2(d+1) weights

$Pr(y=i|x) = \dfrac{\exp(-z_i)}{\exp(-z_1) + \exp(-z_2)}$ * i will either be 1 or 2, so we'll choose 1 to illustrate this example

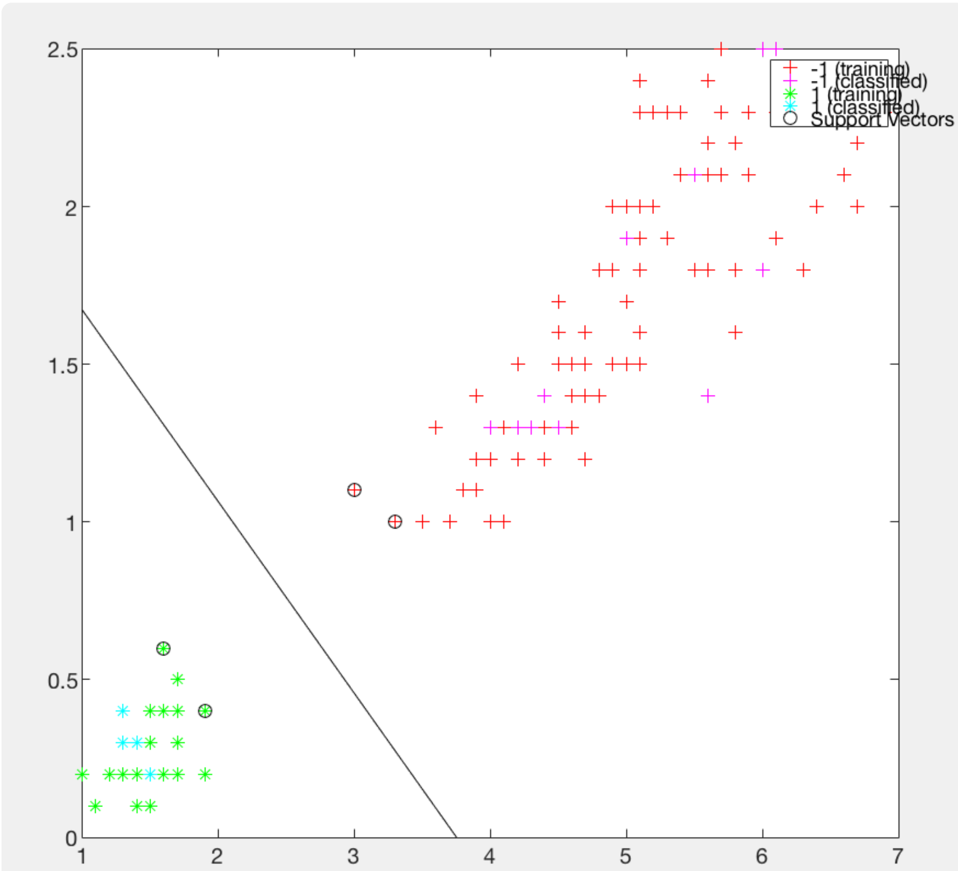
$\Rightarrow Pr(y=1|x) = \dfrac{1}{1+\exp(-(z_1-z_2))}$ & logistic $= Pr(y=1|x) = \dfrac{1}{1+\exp(-(w^T x+w_0))}$

For these two probabilities to be equal, $(z_1-z_2) = \underbrace{(w^T x + w_0)}$ refering to these w now as α

$(z_1-z_2) = (w_1^T x + w_{10}) - (w_2^T x + w_{20})$ (by def of z)

$\alpha^T x + \alpha_0 = w_1^T x - w_2^T x + w_{10} - w_{20} \Rightarrow \left.\begin{array}{l} \alpha = w_1 - w_2 \\ \alpha_0 = w_{10} - w_{20} \end{array}\right\}$ proves $2(d+1)$ for softmax (w) can be reduced to $d+1$ for regression (α)

3)a) Below is a plot of both training data and test data.



Error rate: 0%

3)b) Error rates for types of kernels:

linear: 0%
polynomial: 6.67%
gaussian radial basis: 0%

3)c) Error rates for types of kernels in kernel perceptron:

linear: 66.67%
polynomial: 26.67%
gaussian radial basis: 26.67%

3)d) Error rates for discriminative learning and generative learning:
generative: 33.33%
linear: 20%
polynomial: 6.67%
gaussian radial basis: 6.67%