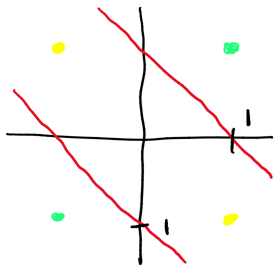
Problem Set 6

Friday, April 16, 2021 10:00 AM

1) a)



class −1

class 1

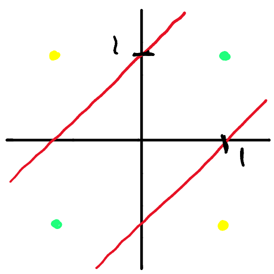
graphed $(x+y)^2 < 2$

$z = (x_1 + x_2)^2$ either 0 or 4, so max margin classifier is 2

if $z < 2$, then 1

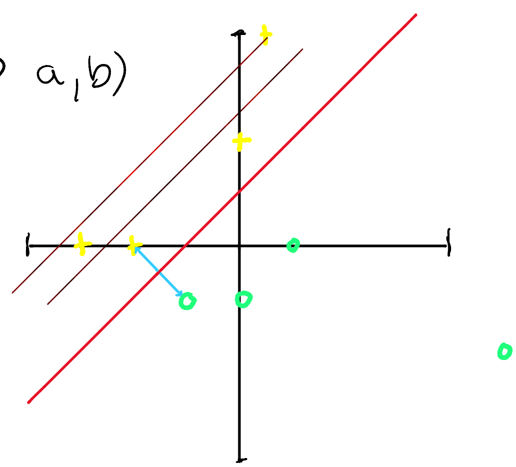
if $z > 2$, then −1

b) x_3 : 0, 2, 2, 0 for respective x_1 & x_2



max margin classifier ⇒ 1

2) a, b)



c) eqn of maximal classifier

$y = x + 1$

eqn of margins

$y = x$, $y = x + 2$

d) if $x_1 - x_2 + 1 > 0$ classify to circle

else classify to cross

e) Left margin

$\beta_0 = 2$

slope $= 2/3$

$x_2 = \frac{2}{3}x_1 + 2$

right margin

$\beta_0 = -\frac{1}{3}$

slope $= 2/3$

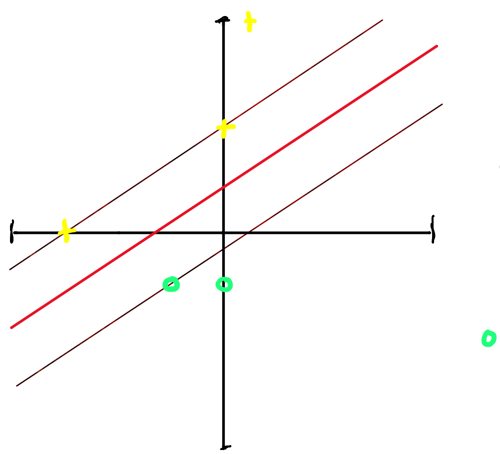
$x_2 = \frac{2}{3}x_1 - \frac{1}{3}$

decision boundary

$\beta_0 = \frac{2 - \frac{1}{3}}{2} = \frac{5}{6}$ $\qquad x_2 = \frac{2}{3}x_1 + \frac{5}{6}$

maximal margin classifier

$f(x) = \frac{2}{3}x_1 - x_2 + \frac{5}{6}$



eqn of separating hyperplane does change

f) $f(x) = \frac{2}{3}x_1 - x_2 + \frac{5}{6} = 0$

$f(x) > 0 \implies$ cross

$f(x) < 0 \implies$ circle

points: $(0,2)\ (-3,0)(-1,-1)$

$f(x) = \sum_{i \in S} \alpha_i \langle x_1, x_2 \rangle + \beta_0$

$= \alpha_1 (x_1, x_2)\begin{bmatrix} 0 \\ 2 \end{bmatrix} + \alpha_2 (x_1, x_2)\begin{bmatrix} -3 \\ 0 \end{bmatrix} + \alpha_3 (x_1, x_2)\begin{bmatrix} -1 \\ -1 \end{bmatrix} + \beta_0$

$= 2\alpha_1 x_2 + (-3)\alpha_2 x_1 + \alpha_3(-x_1 - x_2) + \beta_0$

$= 2\alpha_1 x_2 - 3\alpha_2 x_1 - \alpha_3 x_1 - \alpha_3 x_2 + \beta_0$

$= (2\alpha_1 - \alpha_3)x_2 + (-3\alpha_2 - \alpha_3)x_1 + \beta_0$

$2\alpha_1 - \alpha_3 = -1$ $\quad a_1 = \frac{1}{2} \quad a_2 = -\frac{8}{9} \quad a_3 = 2$

$-3\alpha_2 - a_3 = 2/3$

3) a) 3
- linear boundary
- small margin, so has few support vectors

b) 4
- linear boundary
- large margin, so more support vectors

c) 5
- kernel has linear & squared terms
- has polynomial bounds

d) 6
- radial kernel
- Larger gamma means more overfit decision boundary

e) 1
- radial kernel
- smaller gamma means smoother decision boundary due to larger polling range

4) The gamma parameter determines how much influence each point has on the support vector classifier. If the value is low, then points very far from the classifier are used in calculating the classifier (in addition to closer ones) which will result in a smoother line w/ maybe more misclassifications (higher bias, lower variance). If the gamma value is high, then only the pts. closest to the classifier will have significant impact on the classifier, which will result in a much more wavy classifier that fits the boundary pts. more closely w/ fewer misclassifications (lower bias, higher variance). The gamma controls the

misclassifications (lower bias, higher variance). The gamma controls the granularity b/c it controls the smoothness of the support vector classifier.

5) a) $f(x) = \sum_{i=1}^{N} y_i \lambda_i \exp(-\gamma \|x_i - x_j\|^2) + \beta_0$

b) plug in $\lambda_i = 1$ for all i, $\beta_0 = 0$

$$f(x) = \sum_{i=1}^{N} y_i \exp(-\gamma \|x_i - x_j\|^2)$$

$$|f(x_i) - y_j|$$

$$\left| \sum_{i=1}^{N} y_i \exp(-\gamma \|x_i - x_j\|^2) - y_j \right| < 1$$

then we can split summation to disclude where $i=j$, since $i=j$ means that the exp term will resolve to 1

$$\left| \sum_{\substack{i=1 \\ i \neq j}}^{N} y_i \exp(-\gamma \|x_i - x_j\|^2) \cancel{- y_j + y_j} \right|$$

where $i=j$

now that has to be less than the summation of the absolute values, which is:

$$\sum_{\substack{i=1 \\ i \neq j}}^{N} \left| \underline{y_i} \exp(-\gamma \|x_i - x_j\|^2) \right|$$

we know that the y_i term is irrelevant because it has to be either 1 or -1, & when we take the absolute value, the negative will have no impact

$$\sum^{N} \left| \exp(-\gamma \|x_i - x_j\|^2) \right| < 1$$

$$\sum_{\substack{i=1 \\ i\neq j}}^{N} \left| \exp(-\gamma \underbrace{\|x_i - x_j\|}^2) \right| < 1$$

we can replace the distance term w/ the worst case value ε, which will make all of our summation terms uniform, & allow our summation to become a multiplication

$$\sum_{\substack{i=1 \\ i\neq j}}^{N} \left| \exp(-\gamma \varepsilon^2) \right| < 1$$

$$\Rightarrow (N-1)\left|e^{-\gamma\varepsilon^2}\right| < 1$$

we can also remove the absolute value signs because we know that the e term will always be positive. After this, we can just simplify

$$(N-1)\, e^{-\gamma\varepsilon^2} < 1$$

$$e^{-\gamma\varepsilon^2} < \frac{1}{(N-1)} \Rightarrow -\gamma\varepsilon^2 < \ln\left(\frac{1}{N-1}\right)$$

$$\Rightarrow \gamma < \frac{\ln\left(\frac{1}{N-1}\right)}{\varepsilon^2}$$

5 c) We can train an SVM this way, but as we've discussed many times in this class, a perfect performance on the training set, which this would create, is not an optimal way to train a model. A low bias might result in extremely high variance as a result of overfitting to the training data.