• 1. a) $FPR = 0.2 = \frac{FP}{N=30}$ → 30 misclassified

~~$FNR = 1 - 0.2 = 0.8$~~

$\Rightarrow P = 200 - 30 = 170$

$FP = 30(0.2) = \frac{30}{5} = 6 \Rightarrow TN = 30 - 6 = 24$

$\Rightarrow FN = 50 - 6 = 44$

$\Rightarrow TP = 170 - 44 = 126$

$$\boxed{TPR = \frac{126}{170}}$$

b) $FNR = 1 - TPR$

$= 1 - \frac{126}{170}$

$= \boxed{\frac{44}{170}}$

c) $acc = \frac{TP+TN}{P+N} = \frac{126+24}{200} = \frac{150}{200} = \boxed{\frac{3}{4}}$

d) $F1 = \frac{2 \cdot prec \cdot recall}{prec + recall}$

$= \boxed{\frac{2\left(\frac{126}{132} \cdot \frac{126}{170}\right)}{\frac{126}{132} + \frac{126}{170}}}$

$prec = \frac{TP}{P} = \frac{126}{TP+FP}$

$= \frac{126}{126+6} = \frac{126}{132}$

$recall = TPR = \frac{126}{170}$

2.

$$X = \begin{pmatrix} 2 & 3 \\ -1 & 4 \\ 3 & 0 \\ 2 & -1 \end{pmatrix} \qquad X^T = \begin{pmatrix} 2 & -1 & 3 & 2 \\ 3 & 4 & 0 & -1 \end{pmatrix}$$

Gram

$$XX^T = \begin{pmatrix} 13 & 10 & 6 & 1 \\ 10 & 17 & -3 & -6 \\ 6 & -3 & 9 & 6 \\ 1 & -6 & 6 & 5 \end{pmatrix}$$

3. Training data

X is 100×5
row column
where each row is a data point with 5 features

y is 100×1 corresponding to the labels of each data point

w is a 5×1 vector representing the decision boundary (if we make the data zero mean)

ε is the residual or error in our solution
100×1 vector

4. Sensitivity to outliers means our decision boundary will skew towards outliers, and likely not find a "perfect" boundary for linearly seperable data.

---

we can either try to penalize points with large margins

use SVM w/ soft margin SVM (in case of non seperable data)

or use perceptron (potentially with a limit on the number of iterations if the data is not linearly seperable

5. We may want the cluster exemplars to be data points from the training data. The medoid is the point with the minimal avg dissimilarity and is by design guaranteed to be a training point

6. Soft margin SVM works when the data is not linearly seperable. It introduces a slack variable for each data point allowing for margin errors

7. margin

$$z(x) = \frac{y(w^T x - t)}{\|w\|} \qquad w = (3, 2, -1) \quad t = 4$$

$$\|w\| = \sqrt{9+4+1} = \sqrt{14}$$

a) $x = (0,0,0) \quad y_i = -1$

$$z(x) = \frac{-1\,(0-4)}{\sqrt{14}} = \boxed{\frac{4}{\sqrt{14}}}$$

b) $x = (2,1,1) \quad y_i = -1$

$$z(x) = \frac{-1\,(6+2-1-4)}{\sqrt{14}} = \frac{-3}{\sqrt{14}}$$ misclassified

c) $x = (1,2,3) \quad y_i = 1$

$$z(x) = \frac{(3+4-3-4)}{\sqrt{14}} = 0$$ on the boundary

8. minimize within-class variance
maximize between-class variance

9. a) H4

we choose hypotheses that maximize (ML) the likelihood ($P(D|H)$)

b) chose H that maximizes a posterior ($P(H|D)$)

$$P(H|D) = \frac{P(H)\,P(D|H)}{P(D) = 0.37}$$

this is H3

• 10. Chebyshev distance

$$L_\infty(x,y) = \|x-y\|_\infty = \max_i |x_i - y_i|$$

points

(1,2), (1,1), (2,1), (3,2), (4,2), (4,3)

dist

(2,3): 1, 2, 2, 1, 2, 2

(4,1.5): 3, 3, 2, 1, 0.5, 1.5

iter 0

cluster 1 · mean: (2,3) : data pts : (1,2), (1,1), (2,1), (3,2) ← til

cluster 2 : pts : (4,2), (4,3)

iter 1

cluster 1 mean

$\frac{(6,6)}{4} = (1.5, 1.5)$ pts. (1,2), (1,1) (2,1) (3,2)

cluster 2 mean

$\frac{(8,5)}{2} = (4, 2.5)$: pts (4,2) (4,3)

dist.

(1.5,1.5) 0.5 0.5 0.5 0.5 1.5 1.5

(4,2.5) 3 3 2 1 0.5 0.5

(4,2) (4,3)

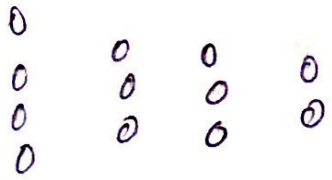
iter 2

same as iter 1

11. we propogate error backwards inorder to update weight connecting nodes, ultimately these weights are what gets output.

12. 3 hidden layers

$4 \times 3 + 3 \times 3 + 3 \times 2 = 12 + 9 + 6 = \boxed{27}$ weights



13. run multiple times with different initial weights and pick the best one

smallest error on validation set

14. $(1.5, 1.0, 1.6) \cdot (0.5, -0.3, 0.2)$

$0.75 + (-0.6) + 0.2 = 0.25$

squared error

$(0.5 - 0.25)^2 = \boxed{(0.25)^2}$

15. Intrinsic dimensionality is 4

16. to transfer ~~data~~ into a higher dimensional space that is linearly separable

17. ℓ could appear $\left\lfloor \frac{|D|}{N} \right\rfloor$ times

18. 2 errors / 5 = 0.4

$\varepsilon = 0.4$ (I realize this isn't quite right)

new weights

$$\left\{ \underbrace{\frac{0.2}{2\varepsilon}, \frac{0.1}{2\varepsilon}}_{\text{increase}}, \underbrace{\frac{0.35}{2(1-\varepsilon)}, \frac{0.3}{2(1-\varepsilon)}, \frac{0.05}{2(1-\varepsilon)}}_{\text{decrease}} \right\}$$

1000/10 = 100

19. each training set has
900 data pts
any given data points d_i will appear in the test set exactly once

20. convolution layer has
$32 \times 11 \times 11 \times 3$ weights to learn
pooling layer
$2 \times 2 \times 2$
+ 54×54 activation maps
+ $64 \times 64 \times 3$ input size