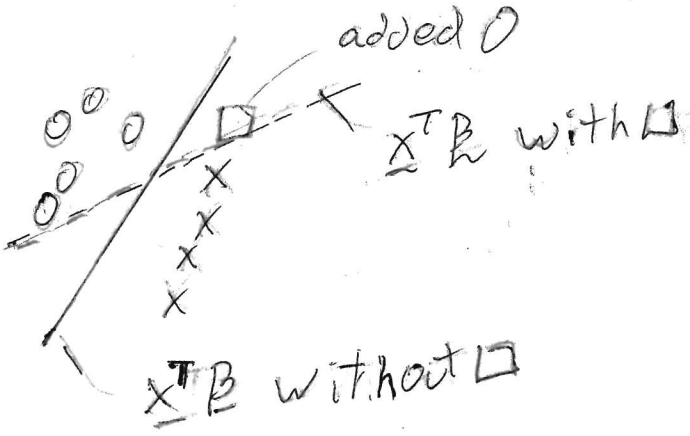


28] SVM maximizes M = width of margin

$\beta_0, \ldots, \beta_p, \varepsilon_1, \ldots, \varepsilon_n$

subject to $\sum_{j=1}^{p} \beta_j^2 = 1$, $\varepsilon_i \geq 0$, $\sum_{i=1}^{n} \varepsilon_i \leq D$

omit β_0 from constraint, $\|\underline{\beta}\|_2^2 = 1$

and $z_i(\beta_0 + \underline{\beta}^T \underline{x}_i) \geq M(1-\varepsilon_i)$,

compare 20].

29] A slack variable $\varepsilon_i = 0$ if $\underline{x}_i$ is on the correct side of the margin. If $\varepsilon_i > 0$, $\underline{x}_i$ is on the wrong side of the hyperplane.

So $z_i(\beta_0 + \underline{\beta}^T \underline{x}_i) \geq m$ has $\varepsilon_i = 0$ and is necessary for $\underline{x}_i$ to be on the correct side of the margin. If $z_i(\beta_0 + \underline{\beta}^T \underline{x}_i) \geq M(1-\varepsilon_i)$ with $\varepsilon_i > 0$ (but not if $\varepsilon_i = 0$), then

$\underline{x}_i$ is on the wrong side of the hyperplane (misclassified). See 17].

30} Let the kernel function be $K(\underline{x}_i, \underline{x}_j)$.

A linear kernel is $K(\underline{x}_i, \underline{x}_j) = \underline{x}_i^T \underline{x}_j$.

A polynomial kernel of degree d is

$K(\underline{x}_i, \underline{x}_j) = [1 + \underline{x}_i^T \underline{x}_j]^d$.

A radial kernel is $K(\underline{x}_i, \underline{x}_j) =$

$$\exp\left[-\gamma \sum_{k=1}^{p} (x_{ik} - x_{jk})^2\right] = \exp\left[-\gamma \| \underline{x}_i - \underline{x}_j \|_2^2\right].$$

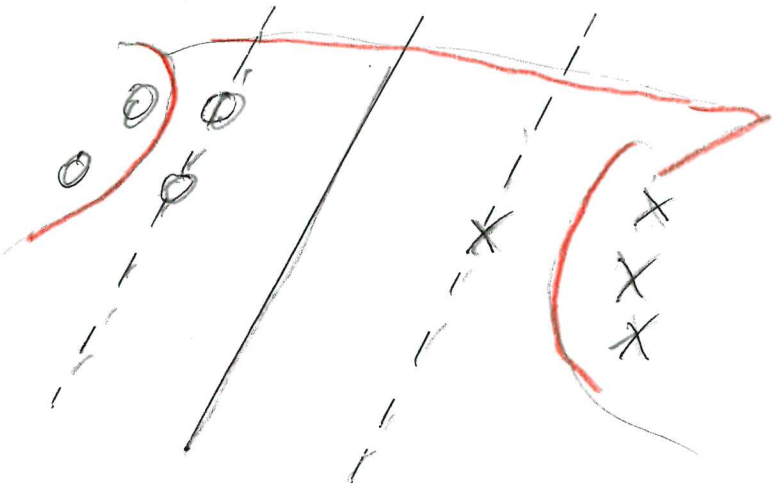
31} A SVM uses $sp = f(\underline{x}) =$

$$\beta_0 + \sum_{i=1}^{n} \alpha_i K(\underline{x}, \underline{x}_i) = \beta_0 + \sum_{i \in S} \alpha_i K(\underline{x}, \underline{x}_i)$$

$= sp = sp(\underline{x})$ where S is the index set of support vectors.

Note: the support vectors (boundaries) determine the hyperplane and margin: if they are moved, the hyperplane moves, too.

See sketch above 28).

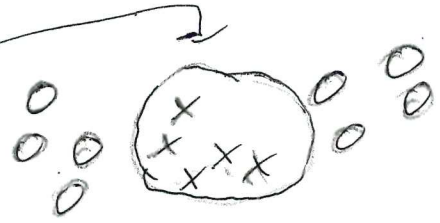


these points are not support vectors.

32] Using $K(\underline{x}, \underline{x}_j)$ leads to nonlinear decision boundaries if K is nonlinear.
The kernel is a bivariate transformation.
There are $\binom{n}{2} = \frac{n(n-1)}{2}$ distinct pairs $(\underline{x}_i, \underline{x}_j)$ that are needed to estimate β_0 and the α_i.

An SVM with $\hat{f}(\underline{x}) = \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\alpha}_i K(\underline{x}, \underline{x}_i)$
$= ESP = ESP(\underline{x})$ is a competitor for QDA
while the SVM with $\hat{f}(x) = \hat{\beta}_0 + \underline{\beta}^T \underline{x}$ is
a competitor for LDA.

ex] boundaries



radial kernel

33] If $\underline{x}$ is far from the $\underline{x}_i$, then

$\|\underline{x} - \underline{x}_i\|_2^2$ is large so

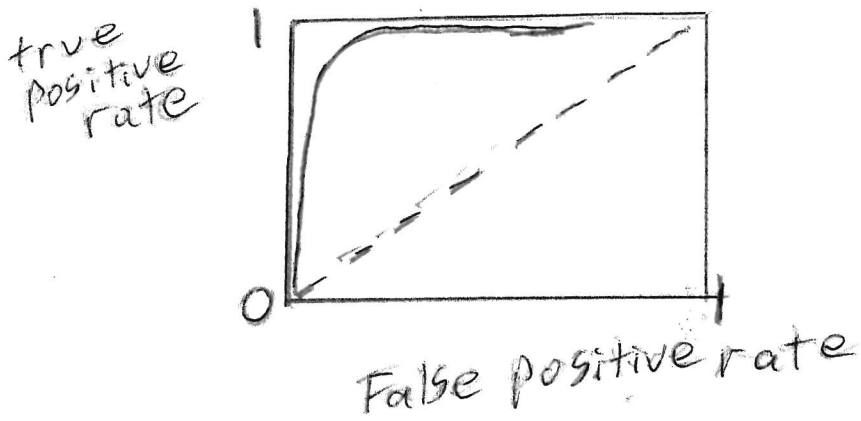
$K(\underline{x}, \underline{x}_i) = \exp(-\gamma \|\underline{x} - \underline{x}_i\|_2^2)$ is tiny, and $\underline{x}_i$ has almost no contribution in $f(\underline{x})$.

(Analogy KNN.)

(← from quality control)

34] A (receiver operating characteristic) ROC curve is used to evaluate binary classifiers. The overall performance is summarized by the area under the ROC curve (AUC). An ideal ROC curve is close to the top left corner of the plot, so the larger the AUC, the better the classifier.

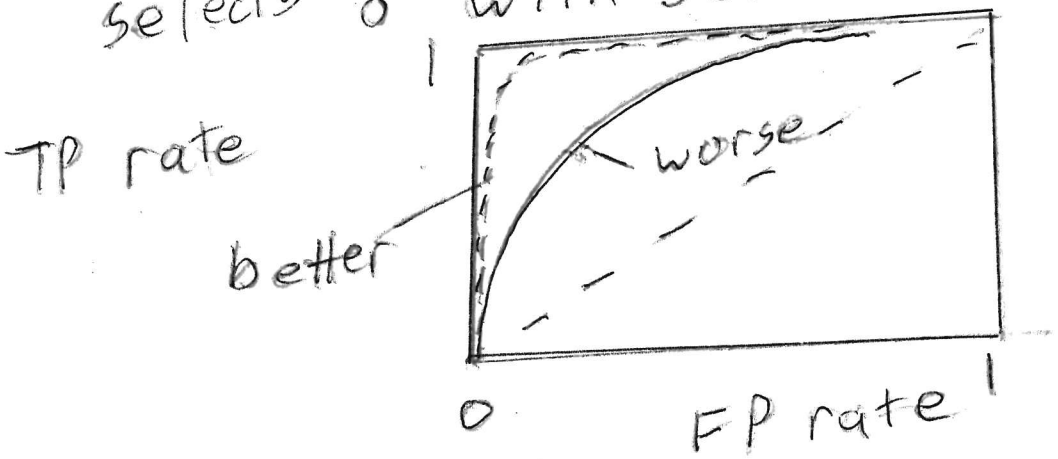
predict \ truth	−1	1	total
−1	true negative TN	false negative FN	N*
1	FP false positive	TP true positive	P*
total	N	P	



$0 \leq AUC \leq 1$

A classifier with AUC = 0.5 does no better than chance.

35] Plot varies γ for radial kernel and selects γ with best ROC curve



36] ROC from test data or validation data is better than ROC from training data.

37] The true positive rate is called the sensitivity and the false positive rate is 1 − specificity.

False positive rate $= \frac{FP}{N} \approx$ type I error, 1 − specificity

True positive rate $= \frac{TP}{N} \approx$ 1 − type II error, power, sensitivity, recall

positive predicted value $\frac{TP}{P^*}$ = precision = 1 − false discovery proportion

negative predicted value $= \frac{TN}{N}$.

38] Sometimes one error is much more important than the other.

ex] loan default versus do not default
Misclassifying "default" should be small.
Misclassifying "do not default" is not as important.

More than $G > 2$ classes from binary classifiers: (ex $m = \binom{3}{2} = 6$: m_1, m_2, m_3)

39] The one versus one or all pairs classifier constructs $\binom{G}{2}$ binary classifiers, one for each distinct pair of groups.
Classify $\underline{x}$ with $f_{ij}(\underline{x})$ and let (2 groups i and j)
m_k = # times $\underline{x}$ is predicted to be in class k.
Then $\hat{y}(\underline{x}) = d$ where $m_d = \max(m_1, \ldots, m_G)$.

40] The one versus all classifier fits G binary SVMs (classifiers; eg); group $i = 1$ versus the other $G-1$ classes with -1

denoted by $\hat{f}_i(\underline{x})$. Then

$\hat{y}(\underline{x}) = d$ where $\hat{f}_d(\underline{x}) = \max\left(\hat{f}_1(\underline{x}), \ldots, \hat{f}_G(\underline{x})\right)$.

ESPs

41] Rules 39] and 40] can be applied to any binary classifier, eg logistic regression.

classification trees and related methods

1] Classification trees are worse than several alternatives, but are building blocks for a good HD classifier.

SUMS

2] A classification tree is a "flexible" method for classification that produces a graph called a tree. Each branch has a label like $\underbrace{x_i > 7.56}_{x_i \text{ quantitative}}$

or $\underbrace{x_j = a, c}_{x_j \text{ a factor (categorical) taking on levels } a, b, c, d, e, f, \text{ say.}}$

Unless told otherwise, if the condition is true, go to the left of the branch. Otherwise, go to the right of the branch. (Some software switches this. Check story problem.)

The bottom of the tree has labels that give $\hat{y}$ for $\underline{x}$. $\hat{y}$ is a label for the classification group.

For example, if $y = \begin{cases} F=0 \\ M=1 \end{cases}$, we could have $\hat{y} \in \{0,1\}$ or $\hat{y} \in \{F, M\}$.

3) Know Given a classification tree and $\underline{x}$ values, find $\hat{y}$. See HW 11.

ex) (Berk 2008 p145-7) Prison data

$\hat{y}$ = no misconduct, minor misconduct, or major misconduct

$\hat{y}$ = 0, 1, 2

HD 90

x_1 = term = sentence length in years

x_2 = agerec (enter prison at)

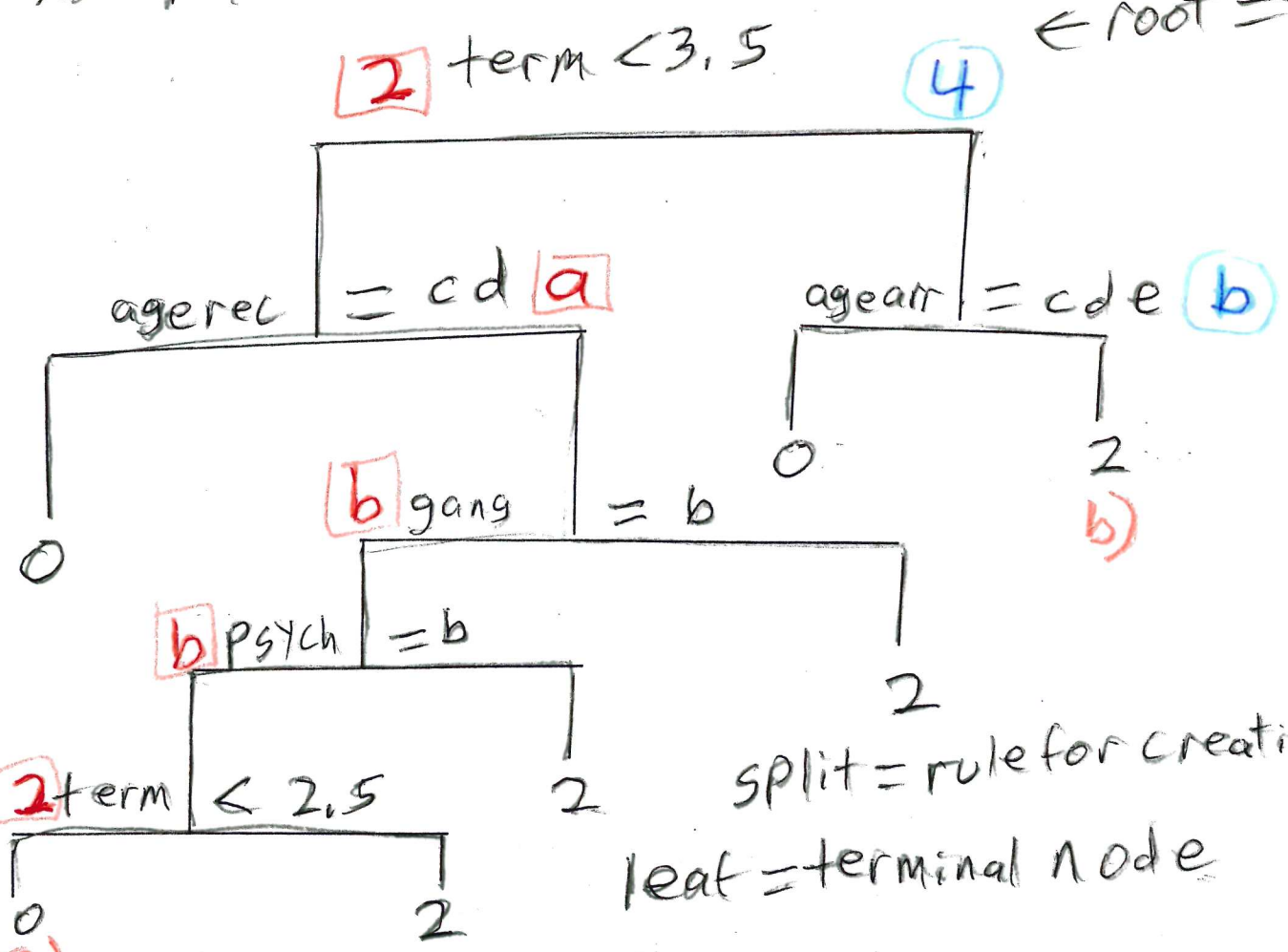
a: 16-20, b: 21-26, c: 27-35, d: 30-35, e: ≥ 33

x_3 = agearr (1st arrest) a: 0-17, b: 18-21,

c: 22-29, d: 30-35, e: ≥ 36

x_4 = gang a = gang activity, b = no gang activity

x_5 = psych a = mental illness b = no mental illness



a) classify if x_1 = term = 2, x_2 = agerec = a,

x_4 = gang = b, x_5 = psych = b

From the red square values, $\boxed{\hat{y} = 0}$.

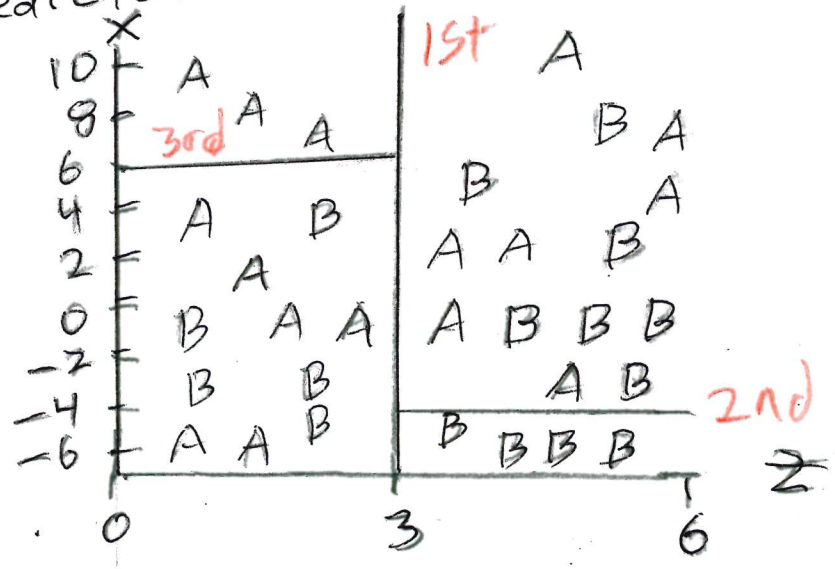
b) Classify if $x_1 = \text{term} = 4$, $x_3 = \text{age arr} = b$

From the blue circle values, $\boxed{\hat{y} = 2}$.

4] Regression trees are similar, but the leaves give $\hat{y} \in \mathbb{R}$ rather than a classification label.

5] Trees that use recursive partitioning for classification and regression trees use the CART algorithm.

ex] $y = A$ or B (classification) with predictors x and z



breaking up x & z is better than breaking up x if y depends on x only through x & z