Problem 1 : more robust against noisy samples.

problem 2: $J(a) = \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1} \sum_{j=1} a_i \cdot a_j \cdot y_i \cdot y_j \cdot \underbrace{\langle x_i, x_j \rangle}_{K(x_i,\, x_j)}$

Similarity measure.

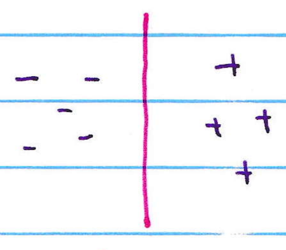
① polynomial K
② RBF K. $= e^{x}$; $x = - \frac{\|x_i - x_j\|_2^2}{2 \cdot \sigma^2}$

Problem 3 : SV = defines the hyperplane that seperates +s from -s
significance = reduced complexity, not relying on non SVs.

Problem 4 : $\min_{\theta} \boxed{C \cdot} \underbrace{\sum y_i \text{cost}_1 (\theta)^T x_i + (1-y_i) \text{cost}_0 (\theta)^T x_i}_{\text{classification error}}] + \frac{1}{2} \cdot \underbrace{\sum_{j=1}^{d} \theta_j^2}_{\text{max margin}}$

if $C = \infty \Rightarrow$ only care about min classification error
$C = \boxed{-\infty} 0 \Rightarrow \frac{1}{2} \sum_{j=1}^{d} \theta_j^2 \Rightarrow$ only care about max the margin.

Problem 5 :



dimensionality of HP = d of samples - 1
HP = decision boundary.

problem 6 : kernel tricks replaces $\langle x_i, x_j \rangle$ in the Lagragian dual with $K(x_i, x_j)$: raises lower-dimensional samples to higher dimensions without explictly doing the mapping.

Problem 7 : $a_j = 0.1 \Rightarrow x_j$ is a support vector.

Problem 8: Primal $= \frac{1}{2}\sum_{j=1}^{d} \theta_j^2$, s.t. $\underbrace{y_i(\theta^T x_i + b) \geq 1}_{\text{Constraints}}$

why? do we want the dual ⇒ the dual internalized the constraints into the objective function.

Problem 9: kenel tricks.

Problem 10: pros: reduced comp times/complexity
achieved linear seperability
cons: assumption that raising to higher dim. is meaningful
preserving data integrity.

Problem 11 what is $H(X,Y)$

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log_2 p(x,y)$$

$$= -\left(\sum_x P(x,0)\log_2 P(x,0) + P(x,1)\log_2 P(x,1)\right)$$

$$= -P(0,0)\log_2 P(0,0) - P(0,1)\log_2 P(0,1) - P(1,0)\log_2 P(1,0) - P(1,1)\log_2 P(1,1)$$

Problem 12. $H(\overset{\text{cloudiness}}{Y} | X=1) = -\sum_{y \in Y} P(y|X=1)\log_2 P(y|X=1)$

$$P(y|x) = \frac{P(x,y)}{P(x)} \qquad = -\sum_{y\in y} \frac{P(X=1,y)}{P(x)} \log_2 \frac{P(X=1,y)}{P(x)}$$

$$= -\frac{P(X=1,y=0)}{P(x)}\log_2 \frac{P(X=1,y=0)}{P(x)} \qquad P(X) =$$

$$-\frac{P(X=1,y=1)}{P(x)}\log_2 \frac{P(X=1,y=1)}{P(x)}$$

Problem 13 $H(Y|X) = \sum_{x \in X} p(x) H(Y|x)$

previously, we have $H(Y|X=1)$; use the same approach to get $H(Y|X=0)$

$$H(Y|X) = P(X=0) H(Y|X=0) + P(X=1) H(Y|X=1)$$
$$= \frac{75}{100} \cdot H(Y|X=0) + \frac{25}{100} \cdot H(Y|X=1)$$

problem 14. $IG(Y|X) = H(Y) - H(Y|X)$

Problem 15. $IG(Y|X)$, when x is useless $\Rightarrow IG(\overset{Y}{\cancel{X}}|X) = 0$

problem 16. $IG(Y|X)$, when X is equivalent to Y) $\Rightarrow IG(Y|X) = H(Y)$

problem 17. Overfitting. $\Rightarrow$ potentially, train acc = 100%, with one leaf node correspond to each one sample.

Problem 18 - Decision Tree Slides 8 & 12.

Problem 19 Decision tree is a greedy heuristics. At each split (down the tree) it picks the feature split thaut guarantees max IG.

problem 20. pro: very interpretable.

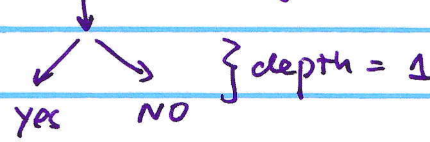
cons: 1. prone to overfit

2. greedy heuristic $\Rightarrow$ tries to find only local optimum

3. time & computation complexity (compared to other non NN approaches)

Problem 21. a weak leaner = a model with > 50% accuracy.

example = decision stumps



Problem 22. misclassified samples get bigger weights

correctly classified " " smaller weights

Problem 23. ① simplicity

② interpretability

Problem 24. $w_{t,i}$. t: iteration / trial #

i = the index of ~~of~~ a sample, x_i.

$w_{t,i} \neq$ (not usually) $w_{t+1,i}$

~ cross entropy loss

Problem 25 $J_{reg_t}(\theta) = -\sum_{i=1}^{n} \boxed{w_{t,i}} \underline{[y_i \log h_\theta(x_i) + (1-y_i)\log(1-h_\theta(x_i))]}$

$+ \boxed{\lambda \|\theta[1:d]\|_2^2}$ → regularization (want to minize θ, to prevent any index from getting too big)

regularized cost for the t th iteration.