SMAI-M20-L23: SVMs and Kernels

C. V. Jawahar

IIIT Hyderabad

October 5, 2020

Announcements

- The number of students who watch and prepare for the sessions is low.
- A number of students are finding it hard to find time. Partly understandable in these times. But not fully acceptable.
- As we move forward, you could expect questions beyond MCQ. (in CR, Quiz, In class). Need to nurture the skill of solving problems on paper.
- Fine tuning the evaluation scheme.

Item	Orig.	Updated
Class Review ($\Uparrow$)	10	25
Home Works ($\Downarrow$)	40	30
Quiz	25	25
Assignments ($\Downarrow$) Problem Solving ($+$)	25	20

Class Review

$$z = \kappa(p, q)$$

We know the kernel $\kappa()$ and the feature map $\phi()$.
Let us start with samples in 2D and

- Understand how $\phi()$ and $\kappa()$ are related in many specific cases.
- Is it unique?

$$\kappa(p, q) = \phi(p)^T \phi(q)$$

Blank

$K_i(\quad)$ $\qquad$ $\phi_i(\quad)$

$$\boxed{K \equiv \sum_{i=1}^{p} K_i(\quad)}$$

$(z_1^T z_2)^2$

$$\sum_{i=\cdot}^{p} (z_1^T z_2)^i$$

$$\sum_{i=1}^{p} (z_1^T z_2)^i$$

Recap:

- **Supervised Learning:** Formulation, Conceptual Issues, Concerns etc.
- **Classifiers:** (i) Nearest Neighbour, (ii) Notion of a Linear Classifier (iii) Perceptrons (iv) Bayesian Optimal Classifier (v) Logistic Regression (vi) Multiclass classification architectures (v) SVMs
- **Dimensionality Reduction and Applications:** (i) Feature Selection and Extraction (ii) PCA (iii) LDA (iv) Eigen face
- **Matrix Factorization and Applications:** (i) SVD, (ii) Eigen Decomposition (iii) Matrix Completion (iv) LSI (v) Recommendations
- **Other Topics:**
  - Linear Regression
  - Probabilistic View, Bayesian View, MLE
  - Gradient Descent: Stochastic and Batch GD
  - Loss Functions and Optimization
  - Eigen Vector based optimization
  - Neuron model, Single Layer Perceptrons
  - Kernel Functions and Kernel Matrix

Lin

$\downarrow$

Non-Lin

This Lecture:

1. Soft Margin SVMs
   - SVM as a classifier that maximizes the margin.
   - How do we make the constraints "soft".
2. Decision Tree Classifier
   - Popular. Simple. Interpretable.
   - Recursive Design. Node test.
3. Kernel Perceptron
   - Illustrative: Kernalizing a linear algorithm.

Questions? Comments?

Blank

$$q_1(p)^T q_1(q) + q_2(p)^T q_2(q) + q_3(p)^T q_3(q)$$

$$\begin{bmatrix} q_1(p) \\ q_2(p) \\ q_3(p) \end{bmatrix} \begin{bmatrix} q_1(q) \\ q_2 \\ \ \end{bmatrix}$$

Blank

Discussions Point - I

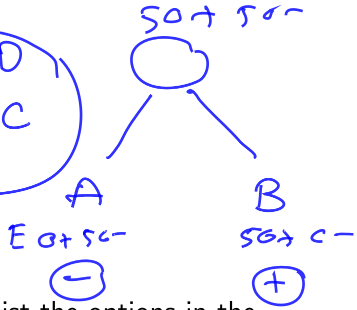
Consider there are 100 samples in 6 dimensions (i.e., $N = 100$ and $d = 6$) and a binary classification (50 each in class + and class -) (i.e., (50+,50-)

If we use ith (1 to 6) feature based the node-test at the root, the two subsets formed are as

- A (25+,25-) and (25+,25-)
- B (45+,35-) and (5+,15-)
- C (25+,2-) and (25+,48-)
- D (40+,10-) and (10+,40-)
- E ~~(0+,50-) and (50+,0-)~~
- F (0+,0-) and (50+,50-)

$C > D$

$D > C$

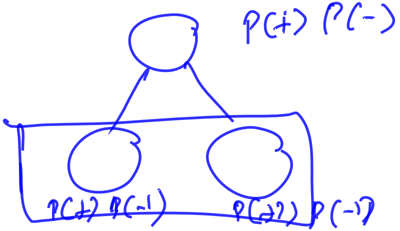


1. What do you prefer as the node-test? (list the options in the decreasing order of your preference)
2. "When $d = 6$, there can be only six possible splits" True or False?
3. Does your decision of node test "guarantee" that this is the best choice? E / A / D > A

Information Gain

In decision tree [1] design, a popular way to do this is byestimating the information gain[2]

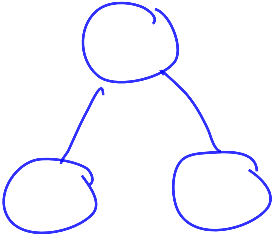
Basic Idea: "Estimate the entropy of the parent set. Estimate the entropy of the children sets (with different attributes) and select the best that removes most uncertainty."



[1] https://en.wikipedia.org/wiki/Decision$_t ree_l earning$

[2] https://en.wikipedia.org/wiki/Information$_g ain_i n_d ecision_t rees$

Blank



$w_1x_1 + w_2x_2 \geq \theta$

"better"

more homogeneous

x

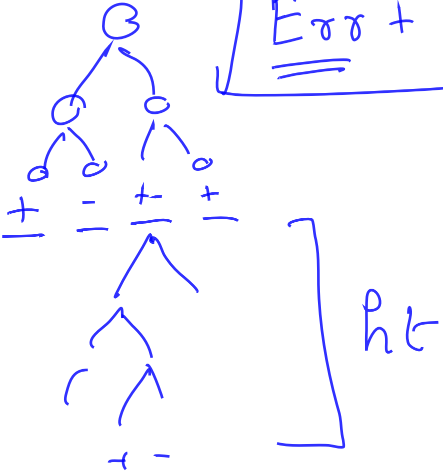
Color =	red
Co =	sv
c =	bl~

Temp > 25

T > 35

Blank

$$\boxed{Err + \lambda\, ht}$$



Blank

Discussion Point - II

We know the softmargin SVM problem as

$$Min \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i^2$$

$\sum_{i=1}^{N} \xi_i$

subject to:

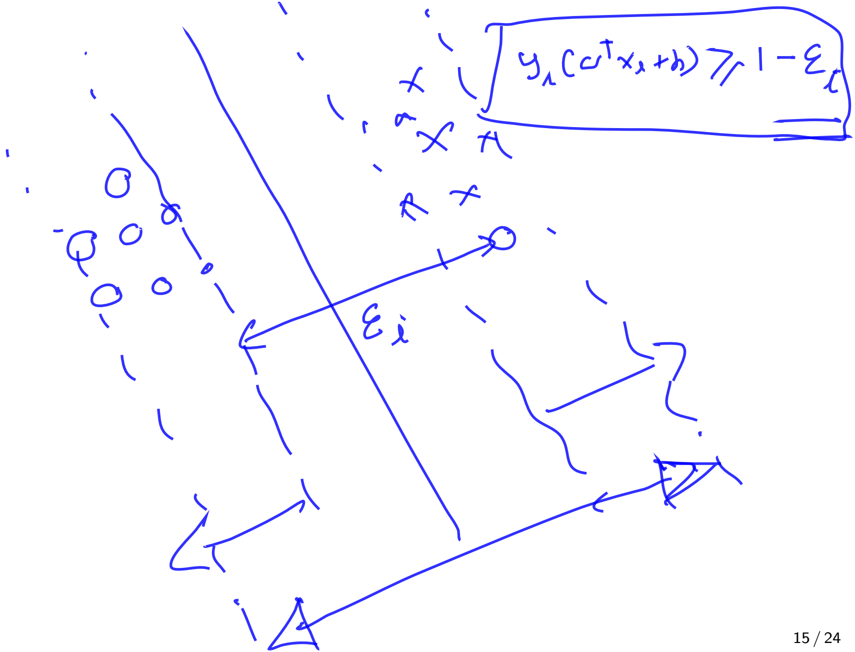
$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \ \ \forall i$$

$$\xi_i \geq 0$$

1. Why $\xi_i \geq 0$? required in the constraint? Why ξ_i in the objective?
2. "If a specific problem has a hard margin possible, ξ_i will all be zero"? True or False?
3. If C is very small (say $C = 0$), what does it mean? what do you expect to see in the final solution?
4. If C is very large (say $+\infty$), what does it mean? what do you expect to see in the final solution?
5. How do we choose C?

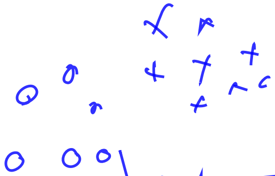
cross-validation

Blank

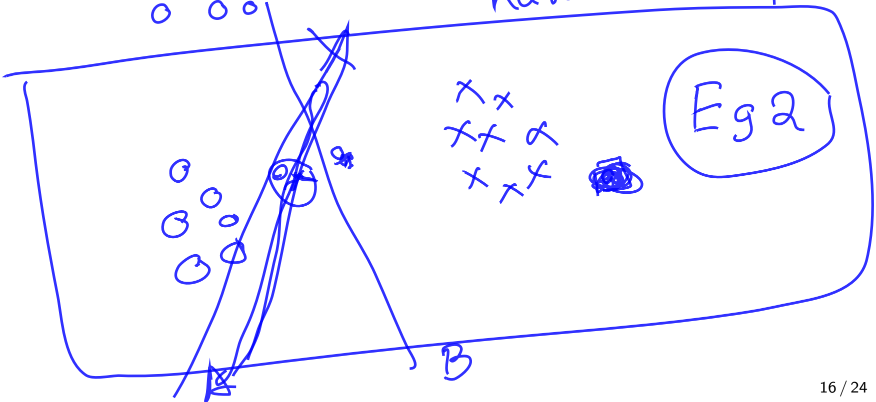


Blank

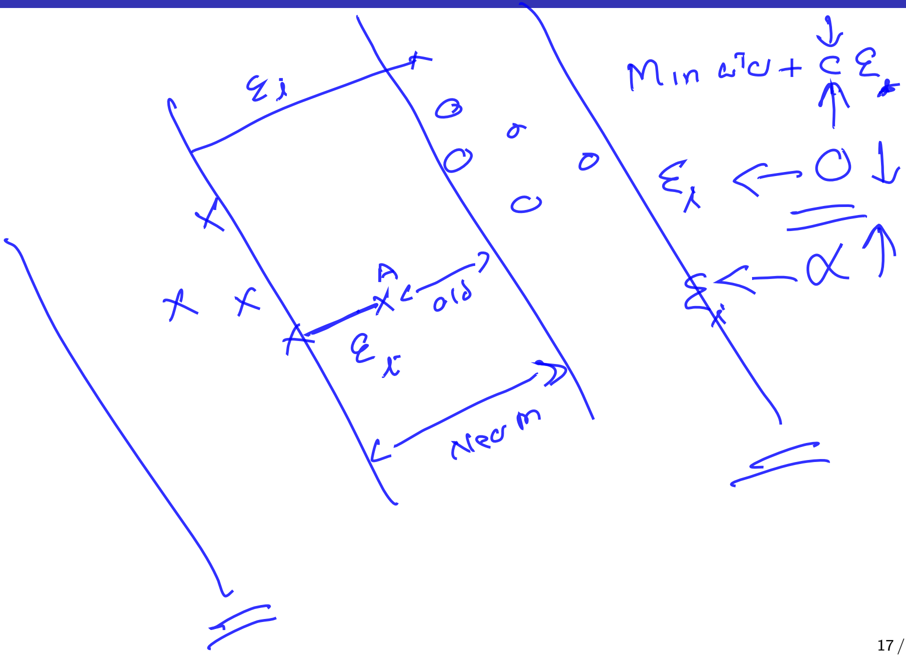
Lin Sep $\Rightarrow \underline{\underline{\varepsilon_i = 0}}$



" hard margin poss!



Blank



Blank

K.PCA

$$w \leftarrow w + y_i x_i$$

$$w = \sum_{i=1}^{N} \alpha_i x_i$$

"Robust"

"express"

Lin Alg (eg. SVM, Perceps)

+

Kernels

= K-ML

Discussions Point -III

We know the perceptron classification as

$$sign(\sum_{i=1}^{N} \alpha_i \mathbf{x}_i^T \mathbf{x}))$$

and the kernel perceptron as

$$sign(\sum_{i=1} \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}))$$

1. Does the kernel perceptron yield a nonlinear boundary?
2. Assume the samples were in 2D, how do we plot (or visualize the decision boundary)?

Blank

Blank

Blank

What Next:?

1. **More on SVMs and Kernels**