Lec 15

Tuesday, November 5, 2019 11:03

Recap: Which linear separator to use?

Idea: maximize the margin

If it's ~~possible~~ possible to linearly separate the data then max margin classifier given by

$$\min_{\beta_0, \beta_1, \dots, \beta_p \in \mathbb{R}} \frac{1}{2}\|\beta\|_2^2$$

$$\text{s.t. } \quad Y_i(\beta^T x_i + \beta_0) \geq 1 \quad \forall i = 1, \dots, n$$

(recall $Y_i \in \{-1, +1\}$)

Nonseparable case Soft margin

Allow some slack $\xi_i \geq 0$

Only going to require

$$Y_i(\beta^T X_i + \beta_0) \geq 1 - \xi_i \quad \forall i$$

But penalize positive slack

We get the

"Support Vector Classifier" (SVC)

$$\min_{\substack{\beta_0, \beta_1, \dots, \beta_p \in \mathbb{R} \\ \xi_1, \dots, \xi_n \in \mathbb{R}}} \frac{1}{2}\|\beta\|_2^2 + C \sum_{i=1}^n \xi_i$$

$$\text{s.t. } \quad Y_i(\beta^T X_i + \beta_0) \geq 1 - \xi_i \quad \forall i = 1, \dots, n$$

$\xi_i \geq 0 \quad \forall i = 1, \dots, n$

Trade off b/w margin ($\|\beta\|_2^2$)
& classification errors ($\sum_i \xi_i$)

Controlled by param $C > 0$

E.g. $C \to \infty \Rightarrow$ get hard margin, max margin classifier (if it's linearly seperable)

Connect & differences to logistic regression

SVC can be equivalently written as:

$$\min \sum_i \max(0, 1 - y_i(\beta^T x_i + \beta_0)) + \frac{1}{2C}\|\beta\|_2^2$$

$$= \min \sum_i loss_{Hinge}(y_i, \beta^T x_i + \beta_0) + \lambda \|\beta\|_2^2$$

$$loss_{Hinge}(y, \hat{y}) = \max(0, 1 - y\hat{y}), \quad \lambda = \frac{1}{2C}$$

Logistic regression w/ ridge regularization

$$\min \sum_i \log\left(1 + e^{-y_i(\beta^T x_i + \beta_0)}\right) + \lambda \|\beta\|_2^2$$

- Both optimize a loss fn
  + (possible) regularization
- Different loss fn
  - Usually, for classification, prefer hinge loss

- for prob prediction, prefer logistic reg.

(SVC gives only a decision boundary, not an estimate $\mathbb{P}(Y=1|X)$)

(to get prob estimates from SVC can use Platt scaling)

Kernelizing the SVM:

What happens if we augment the features x using a feature map:

$$\varphi : \mathbb{R}^p \to \mathbb{R}^q \qquad q \geq p$$

e.g. $\varphi((x)) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$ $\quad p=1,\ q=2$

$\varphi\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \begin{pmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \\ x_1 x_2 \end{pmatrix}$ $\quad p=2,\ q=5$

Get the augmented SVC:

$$\min_{\substack{\beta \in \mathbb{R}^q \\ \beta_0 \in \mathbb{R}}} \frac{1}{2}\|\beta\|_2^2 + C \sum_{i=1}^n \xi_i$$

$$\xi_i \geq 0$$

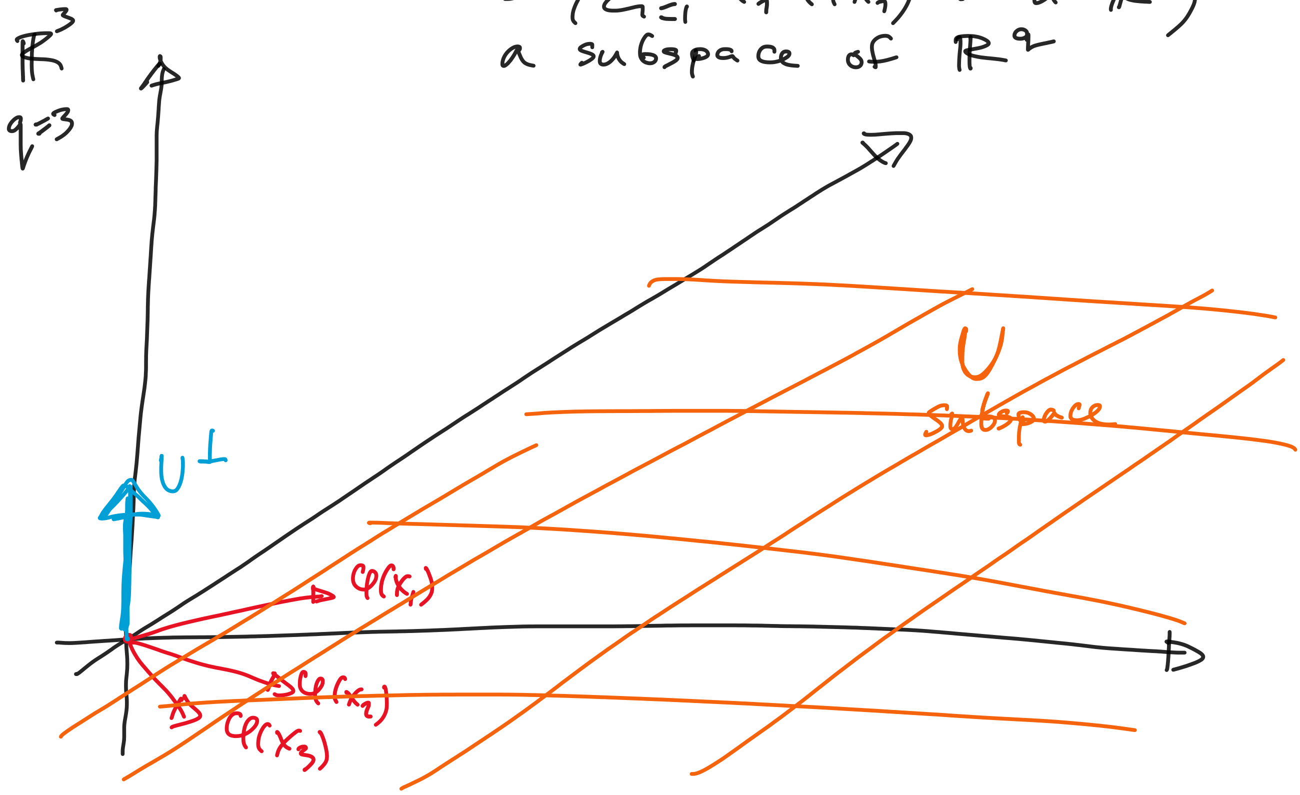
$$\xi_i \geq 1 - y_i(\beta^T \varphi(x_i) + \beta_0)$$

We can even take $q \to \infty$

Consider $U = \text{span}(\{\varphi(x_1), \cdots, \varphi(x_n)\})$

$= \{\sum_{i=1}^{n} \alpha_i \varphi(x_i) : \alpha \in \mathbb{R}^n\}$

a subspace of $\mathbb{R}^q$



Orthogonal complement

$U^\perp = \{v \in \mathbb{R}^q : v^T u = 0 \quad \forall u \in U\}$

$= \{v \in \mathbb{R}^q : v^T \varphi(x_i) = 0 \quad \forall i\}$

Any vector $z \in \mathbb{R}^q$ can be written

as $z = u + v$

where $u \in U$
$v \in U^\perp$

$\|z\|^2 = z^T z = (u+v)^T (u+v)$

$= u^T u + \cancel{2 u^T v}^{\,0} + v^T v$

$= \|u\|^2 + \|v\|^2$

Write $\beta = u + v$ $\quad u \in \mathcal{U} = \text{Span}\{(\varphi(x_1), \ldots, \varphi($
$v \in \mathcal{U}^\perp \quad (v^T \varphi(x_i) = 0 \;\forall$

$$\|\beta\|^2 = \|u\|^2 + \|v\|^2$$

$$\beta^T \varphi(x_i) = u^T \varphi(x_i)$$

Get that the augmented SVC is:

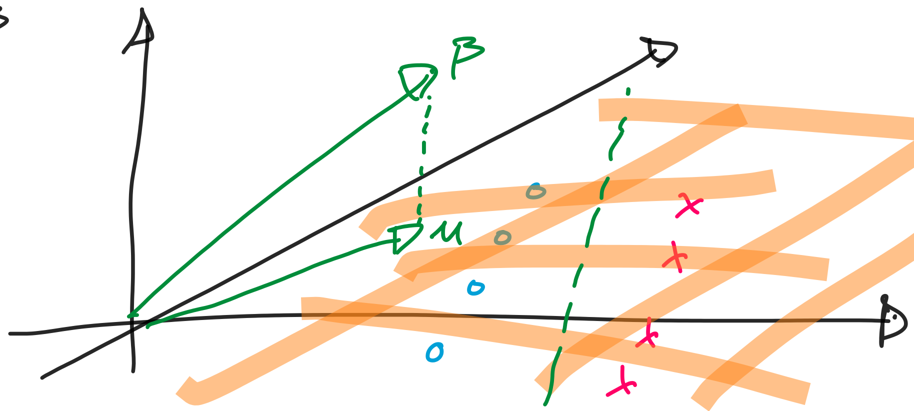
$$\min_{\substack{u \in \mathcal{U},\, \beta_0 \\ v \in \mathcal{U}^\perp}} \frac{1}{2}\|u\|^2 + \frac{1}{2}\|v\|^2 + C \textstyle\sum_i \xi_i$$

$$\xi_i \geq 0$$

$$\xi_i \geq 1 - y_i\left(u^T \varphi(x_i) + \beta_0\right)$$

At optimality get $v = 0$

$\mathbb{R}^3$



Conclusion optimal β can be written as

$$\beta = \textstyle\sum_i \alpha_i \varphi(x_i)$$

(known "representer trick")

Let's do SVC & optimize over these α's inste

$$\min \quad \frac{1}{2}\sum_{ij} \alpha_i \alpha_j \underbrace{\varphi(x_i)^T \varphi(x_j)}_{K(x_i, x_j)} + C \sum_i \xi_i$$

$$\xi_i \geq 0$$

$$\xi_i \geq 1 - \Big(y_i \sum_j \alpha_j \underbrace{\varphi(x_j)^T \varphi(x_i)}_{K(x_i, x_j)} + \beta_0\Big)$$

Define $K(x, x') = \varphi(x)^T \varphi(x')$

$$\underline{K}_{ij} = K(x_i, x_j) \qquad \underline{K} \in \mathbb{R}^{n \times n}$$

Can write augmented SVC as

$$\min_{\substack{\alpha \in \mathbb{R}^n \\ \beta_0 \in \mathbb{R} \\ \xi \in \mathbb{R}^n}} \quad \frac{1}{2} \alpha^T \underline{K} \alpha + C \sum_{i=1}^n \xi_i$$

$$\xi_i \geq 0$$

$$\xi_i \geq 1 - y_i \left(\underline{K}_{i}^T \alpha + \beta_0\right)$$

"kernel trick"