Lecture 12: 2/27/17

Kernels

$D = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N} \quad x^{(i)} \in \mathcal{X}, \; y^{(i)} \in \mathcal{Y}$

Def: function $k(x,z)$ is a **kernel** iff

$\exists \phi : \mathcal{X} \to \mathbb{R}^D$ s.t. $k(x,z) = \phi(x)^T \phi(z)$

Def: An $N \times N$ matrix K is a **kernel matrix** iff

① K is symmetric: $K_{ij} = K_{ji}$

② K is positive-semidefinite: $\forall a \in \mathbb{R}^N, \; a^T K a \geq 0$

(Mercer's Theorem)

Def: For dataset D, kernel k, the **Gram matrix** is

$K_{ij} = k(x^{(i)}, x^{(j)})$

* compute once at the start of training

Note: Can show k is a kernel fn. if for all datasets D the corresponding Gram matrix is symmetric and positive-semidef.

Closure Properties:

Let k_1 and k_2 be kernels.

We can define a new kernel as...

$k(x,z) \triangleq k_1(x,z) + k_2(x,z)$

$k(x,z) \triangleq k_1(x,z)\, k_2(x,z)$

$k(x,z) \triangleq k_1(x,z) + c, \quad c > 0$

$k(x,z) \triangleq c\, k_1(x,z), \quad c > 0$

$k(x,z) \triangleq \exp(k_1(x,z))$

Ex:

$k(x,z) = x^T z$

$k(x,z) = x^T z + 1$

$k(x,z) = (x^T z + 1)(x^T z + 1)$

$k(x,z) = (x^T z + 1)^d$

Perceptron

Predict:

$$\hat{y} = h(\vec{x}) = \text{sign}(\vec{\theta}^T \vec{x})$$

Update:

$$\vec{\theta} \leftarrow \vec{\theta} + y^{(t)} \vec{x}^{(t)} \quad \text{if } y^{(t)} \neq h(\vec{x}^{(t)})$$

Kernel Perceptron

← Online

Predict:

$$\hat{y} = h(\vec{x}) = \text{sign}\left(\sum_{i=1}^{t} \alpha_i k(\vec{x}^{(i)}, \vec{x}) \right)$$

Update:

$$\alpha_t = \begin{cases} y^{(t)} & \text{if } y^{(t)} \neq h(x^{(t)}) \\ 0 & \text{otherwise} \end{cases}$$

String Kernel

Desired features:

$\vec{x} = [c, a, t]$ $\quad \vec{z} = [s, a, t]$

$$k(x, z) = \phi(x)^T \phi(z)$$

= # of common substrings btwn. $\vec{x}$ and $\vec{z}$

Suffix Trees can compute this.

$$\phi(\vec{x}) = \begin{bmatrix} 1 \\ 0 \\ 1 \\ \vdots \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} a \\ b \\ c \\ \vdots \\ z \\ aa \\ ab \\ \vdots \\ zz \\ aaa \\ aab \\ \vdots \\ zzz \end{matrix}$$

$|\Sigma| = 26$

	Standard Perc.	Kernel Perc.
# parameters	$O(\lvert\Sigma\rvert^K)$	$O(N)$
Naive Runtime	$O(\lvert\Sigma\rvert^K)$	$O(K)$
Sparse Runtime	$O(K)$	

# of examples we've seen.

Background : Quadratic Programming

Unconstrained

$$\hat{\theta} = \underset{\vec{\theta} \in \mathbb{R}^M}{\operatorname{argmin}} \; J(\vec{\theta})$$

Constrained

$$\hat{\theta} = \underset{\vec{\theta} \in \mathbb{R}^M}{\operatorname{argmin}} \; J(\vec{\theta}) \quad \text{s.t.} \quad g(\vec{\theta}) \leq \vec{b}$$

Ex # 1 :

Explicit Form

$$\hat{\vec{x}} = \arg\min_{\vec{x}} \; 2(x_1)^2 + x_1 x_2 + (x_2)^2 + x_1 + x_2$$

$$\text{s.t.} \quad x_1 \leq 1/4$$

$$-\tfrac{1}{2} x_1 - x_2 \leq 1/2$$

$$x_2 \leq 3/4$$

Matrix Form

$$\min_{\vec{x}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 2 & 1/2 \\ 1/2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\text{s.t.} \begin{bmatrix} 1 & 0 \\ -1/2 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq \begin{bmatrix} 1/4 \\ 1/2 \\ 3/4 \end{bmatrix}$$

Component-wise vector inequality

Quadratic Program (Standard Form)

Quadratic objective: $\min_{\vec{x}} \; \frac{1}{2} \vec{x}^T Q \vec{x} + \vec{c}^T \vec{x}$

Linear constraint: $\text{s.t.} \quad A\vec{x} \leq \vec{b}$

$Q \in \mathbb{R}^{M \times M}$ $\quad$ $c \in \mathbb{R}^M$

$A \in \mathbb{R}^{N \times M}$ $\quad$ $b \in \mathbb{R}^N$

N: # of constraints, M: # of parameters

Solvers

- Interior Points
- Conjugate Gradient
- Ellipsoid Method
- Primal-Dual

Special Case:

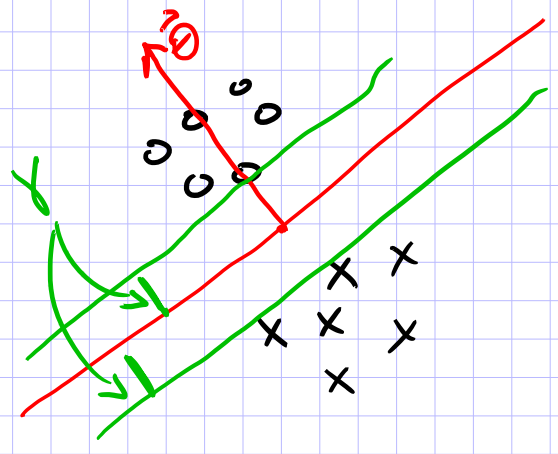
If Q is positive-definite, the problem is convex.

$$q^T Q q > 0, \quad \forall q \in \mathbb{R}^M$$

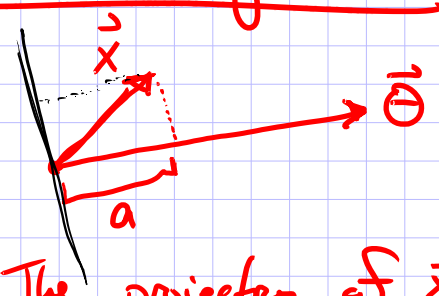
SVM

Recall: Data D is *linearly separable* iff

$$\exists\, \vec{\theta} \in \mathbb{R}^M \text{ s.t. } y^{(i)} \vec{\theta}^T \vec{x}^{(i)} \geq \gamma, \forall i$$
$$\gamma > 0 \qquad \|\theta\|_2 = 1$$



Vector Projection



The projection of $\vec{x}$ onto $\vec{\theta}$ has length

$$a = \frac{\vec{\theta}^T \vec{x}}{\|\vec{\theta}\|_2}$$

Below we "invent" SVMs through a sequence of optimization problem

Key idea: Find $\vec{\theta}$ with maximum margin

- Intively reasonable
- Theoretically well-grounded

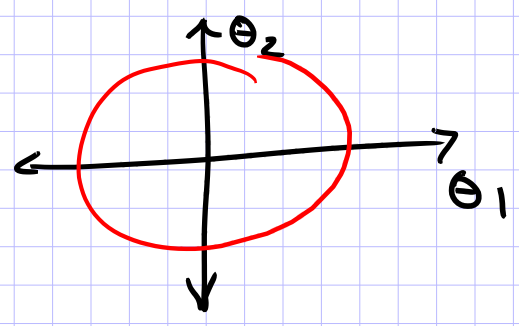
SVM (Linealy Separable case)

Opt #1

Goal: Maximize the margin s.t. all points are correctly classifed

$$\max_{\gamma, \vec{\theta}} \gamma$$
$$\text{s.t. } y^{(i)} \vec{\theta}^T \vec{x}^{(i)} \geq \gamma \quad \forall i$$
$$\|\vec{\theta}\|_2 = 1$$

Problem: $\|\theta\|_2 = 1$ is nonlinear and nonconvex



Opt #2

Define $\vec{w} = \frac{\vec{\Theta}}{\gamma}$

If $\|\vec{\Theta}\|_2 = 1$, then

$$(\|w\|_2)^2 = \frac{1}{\gamma^2}\left(\|\vec{\Theta}\|_2\right)^2 = \frac{1}{\gamma^2}$$

$\Rightarrow$ minimizing $(\|w\|_2)^2$ is equivalent to maximizing γ $\left(\gamma_M \leq \sum_{m=1}^{M} (w_M)^2\right)$

~~$\max_{\gamma,\vec{\Theta},\vec{w}} \gamma$~~ $\min_{\vec{w}} (\|\vec{w}\|_2)^2$

s.t. $y^{(i)} \frac{\vec{\Theta}^T}{\|\vec{\Theta}\|} \vec{x}^{(i)} \geq \gamma \quad \forall i$ $\quad\Rightarrow\quad$ s.t. $y^{(i)} \vec{w}^T x^{(i)} \geq 1, \forall i$

~~$\|\vec{\Theta}\|_2 = 1$~~ irrelevant

$\min_{\vec{w}} (\|w\|_2)^2$

Linear Constraints

Quadratic Objective

★ Standard Form for a Quadratic Program

★ The problem is convex

Problem: not linearly separable?

Opt #3

Goal: Maximize margin and Minimize # of errors

$$\min_{\vec{w}} (\|w\|_2)^2 + C(\#\text{ of errors on train})$$

Problem: NP-Hard to solve this!

Sign() is a step function

Opt #4

Goal: Maximize margin, Minimize Hinge Loss

Def: the **Hinge Loss** is $\max\left(0, 1 - y^{(i)}\vec{w}^T x^{(i)}\right)$

$$\min_{\vec{w}} \left(\|w\|_2\right)^2 + C\left(\sum_{i=1}^{N} \max\left(0, 1 - y^{(i)}\vec{w}^T x^{(i)}\right)\right)$$

$\Updownarrow$ or Equivalently

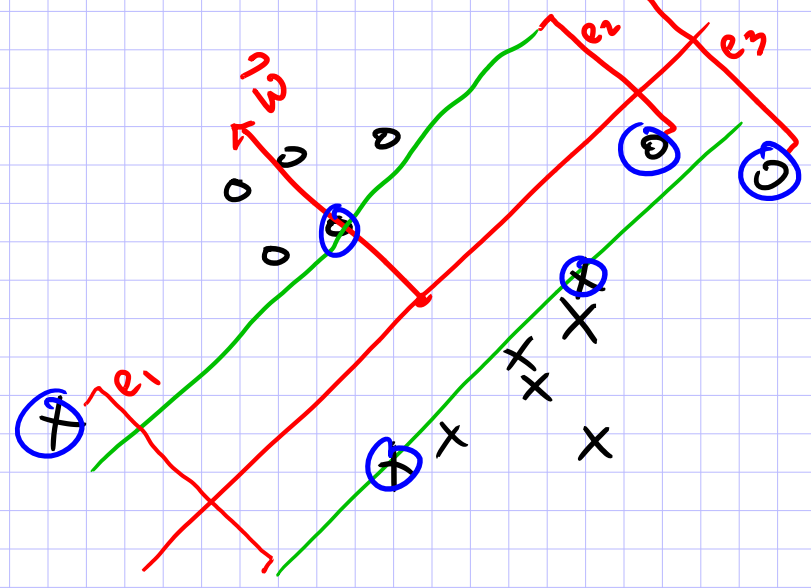
$$\min_{\vec{w},\vec{e}} \left(\|w\|_2\right)^2 + C\left(\sum_{i=1}^{N} e_i\right)$$

$$\text{s.t. } y^{(i)}\vec{w}^T\vec{x}^{(i)} \geq 1 - e_i, \ \forall i$$

$$e_i \geq 0$$

e_i: called "slack variables"

- ★ C is user-chosen hyperparameter which trades off btwn max margin and min error goals.
- ★ Still convex QP solvable by black-box solvers



blue circled points are "support vectors"