LECTURE 11

SPRING 2021

APPLIED MACHINE LEARNING

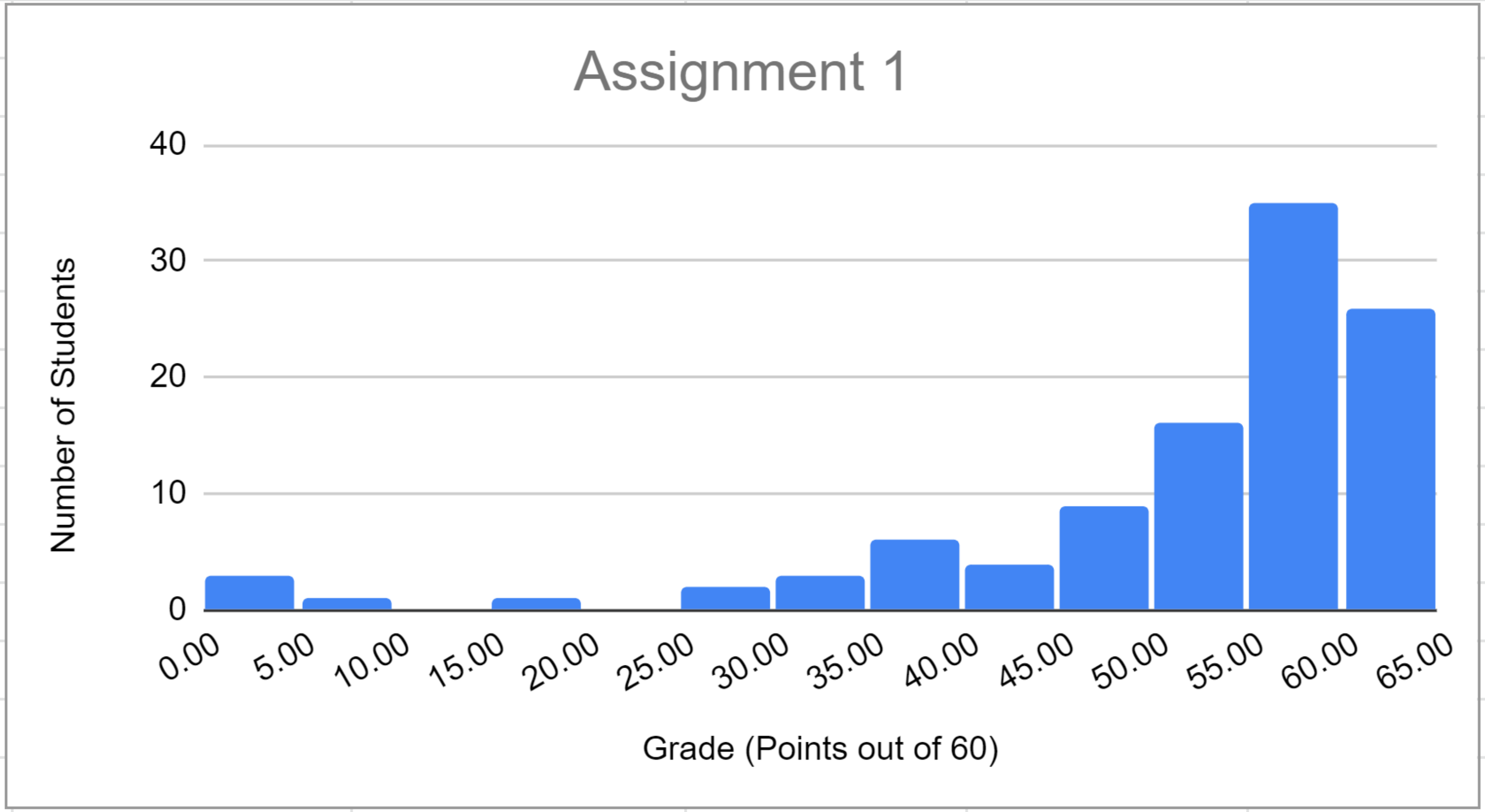
CIHANG XIE

SLIDE CREDIT:

DHRUV BATRA

ANDREW ZISSERMAN

HW1



Average: 50.54
Median: 56.00
STD: 13.28

Q2@HW2

```python
def computeRegularizedCost(X, y, theta, lambdah):
    m = y.size
    J = (np.sum((X @ theta - y)**2))/2/m + lambdah * np.sum([i**2 for i in theta])
    return J


def gradientDescentWithRegularization(X, y, theta, alpha, num_iters, lambdah):
    m = y.shape[0]
    theta = theta.copy()
    J_history = []
    for i in range(num_iters):
        theta_tmp = []
        for j in range(len(theta)): # partial derivative
            gradient = (alpha/m) * np.sum(((X @ theta) - y) * X[:,j]) + 2 * lambdah * theta[j]
            new_theta = theta[j] - gradient
            theta_tmp.append(new_theta)
        theta = theta_tmp
        J_history.append(computeCost(X, y, theta))
    return theta, J_history
```

WEEK 5 GROUP ACTIVITIES

- Some groups have completed the SVM code
- There is one group has not submitted it

TODAY

- Support Vector Machine

  -- review

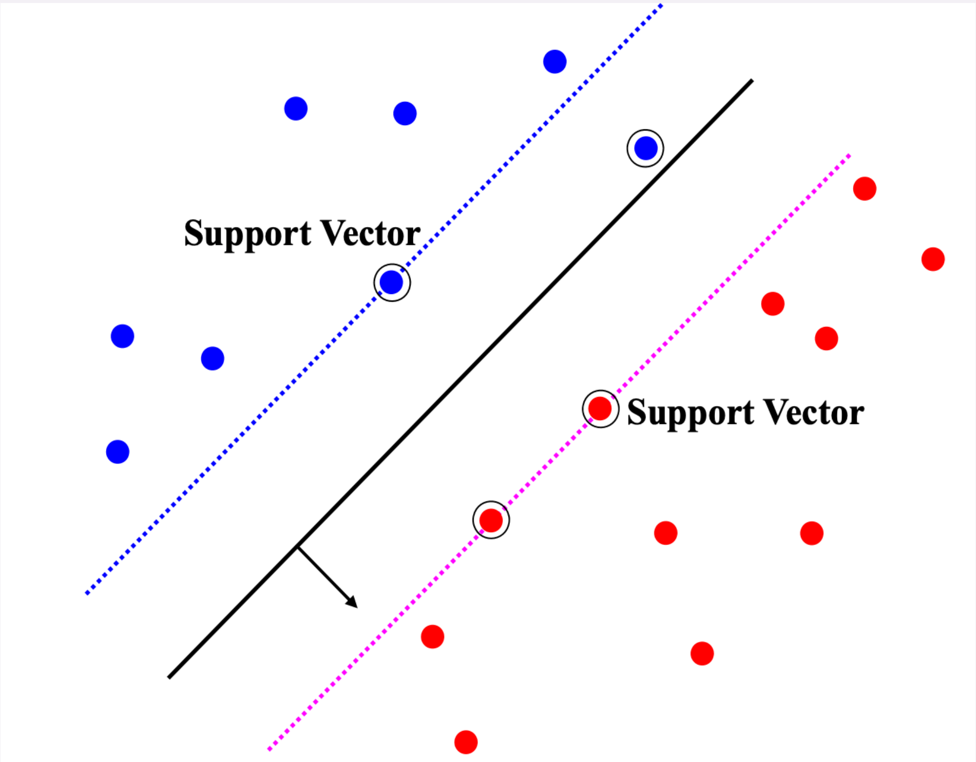
  -- Lagrangian duality

  -- kernel trick

SUPPORT VECTOR MACHINE (SVM)

$$\min_{\boldsymbol{\theta}} \frac{1}{2}\,\boldsymbol{\theta}^T\boldsymbol{\theta}$$

$$\text{s.t.} \quad \boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + \boldsymbol{b}) \geq \mathbf{1}, \forall \boldsymbol{i}$$



SOFT MARGIN SVM

$$\min_{\theta,\xi,b} \frac{1}{2} \theta^T \theta + C \sum_{i=1}^{N} \xi_i$$

$$\text{s.t.} \quad y^{(i)}(\theta^T x^{(i)} + b) + \xi_i \geq 1,$$

$$\xi_i \geq 0, \forall i$$

ξ_i is the "slack" variable

- for $0 < \xi_i \leq 1$ point is between margin and correct side of hyperplane. This is a margin violation
- for $\xi_i > 1$ point is misclassified

SOFT MARGIN SVM

$$\min_{\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{b}} \frac{1}{2} \boldsymbol{\theta}^T \boldsymbol{\theta} + \mathrm{C} \sum_{i=1}^{N} \xi_i$$

$$\text{s.t.} \quad \boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + \boldsymbol{b}) + \xi_i \geq 1,$$

$$\xi_i \geq 0, \forall \boldsymbol{i}$$

C is a regularization parameter:

- small C allows constraints to be easily ignored → large margin
- large C makes constraints hard to ignore → narrow margin
- C = ∞ enforces all constraints: hard margin

GRADIENT DESCENT FOR SVM

$$\boldsymbol{y^{(i)}(\theta^T x^{(i)} + b) + \xi_i \geq 1} \ \& \ \boldsymbol{\xi_i \geq 0}$$

$$\boldsymbol{\xi_i = \max\{0, 1 - y^{(i)}(\theta^T x^{(i)} + b)\}}$$

$$\min_{\boldsymbol{\theta}, \boldsymbol{b}} \frac{\mathbf{1}}{\mathbf{2}} \boldsymbol{\theta^T \theta} + \mathrm{C} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \max\{\mathbf{0}, \mathbf{1} - \boldsymbol{y^{(i)}(\theta^T x^{(i)} + b)}\}$$

GRADIENT DESCENT FOR SVM

$$\text{COST}(\theta, b) = \frac{1}{2}\theta^T\theta + \text{C}\sum_{i=1}^{N} \max\{0, 1 - y^{(i)}(\theta^T x^{(i)} + b)\}$$

$$= \sum_{i=1}^{N}\left(\frac{1}{2N}\theta^T\theta + \text{C}\max\{0, 1 - y^{(i)}(\theta^T x^{(i)} + b)\}\right)$$

For each data point $x^{(i)}$

$$\frac{\partial Cost(\theta, b)}{\partial \theta_j} = \begin{cases} \frac{1}{N}\theta_j - \text{C}\, y^{(i)} x_j^{(i)}, & \text{if } 1 - y^{(i)}(\theta^T x^{(i)} + b) > 0 \\ \frac{1}{N}\theta_j, & \text{otherwise} \end{cases}$$

$$\frac{\partial Cost(\theta, b)}{\partial b} = \begin{cases} -\text{C}\, y^{(i)}, & \text{if } 1 - y^{(i)}(\theta^T x^{(i)} + b) > 0 \\ 0, & \text{otherwise} \end{cases}$$

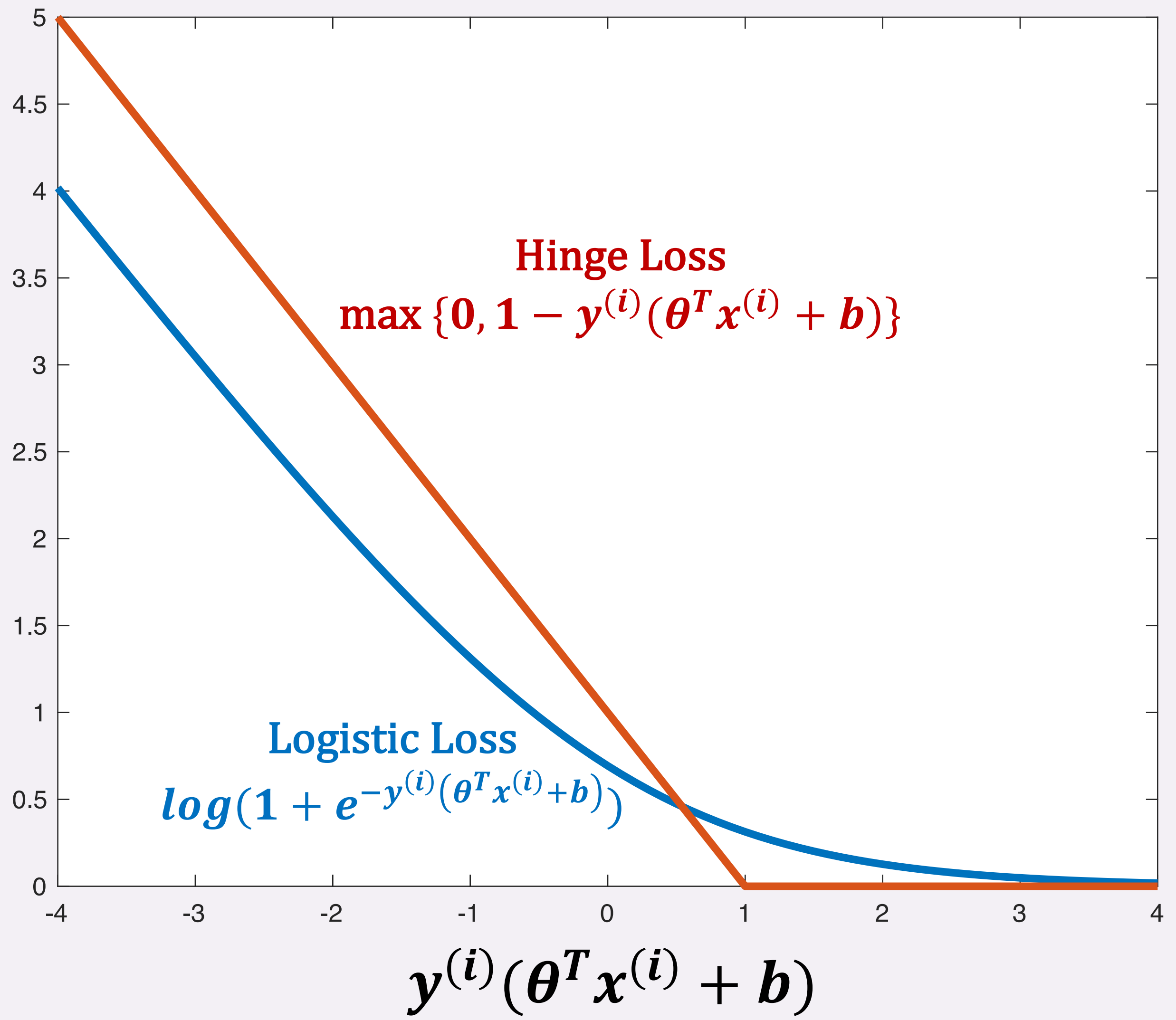
RELATIONSHIP TO LOGISTIC REGRESSION

$$\min_{\theta,b} \lambda\theta^T\theta - \sum_i \log P\left(y^{(i)} \middle| x^{(i)}; \theta, b\right)$$

$$\min_{\theta,b} \underbrace{\lambda\theta^T\theta}_{\text{Regularization}} + \sum_i \underbrace{\log\left(1 + e^{-y^{(i)}\left(\theta^T x^{(i)} + b\right)}\right)}_{\text{Logistics Loss}}$$

$$\min_{\theta,b} \underbrace{\frac{1}{2}\theta^T\theta}_{\text{Regularization}} + \mathrm{C}\sum_{i=1}^{N} \underbrace{\max\left\{0, 1 - y^{(i)}\left(\theta^T x^{(i)} + b\right)\right\}}_{\text{Hinge Loss}}$$

RELATIONSHIP TO LOGISTIC REGRESSION



Logistic loss is sometime viewed as the **smooth version** of the Hinge loss.

DUAL FORMULATION

LAGRANGIAN DUALITY

Primal

$$\min_{\boldsymbol{\theta},\boldsymbol{b}} \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta}$$

$$\text{s.t.} \quad \boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + \boldsymbol{b}) \geq \mathbf{1}, \forall \boldsymbol{i}$$

The Lagrangian for the primal problem:

$\boldsymbol{L} = \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} + \sum_i \boldsymbol{\alpha_i}[\mathbf{1} - \boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + \boldsymbol{b})]$ **where** $\boldsymbol{\alpha_i} \geq \mathbf{0}$ are Lagrange multipliers.

Now we want to solve: $\min_{\boldsymbol{\theta},\boldsymbol{b}} \max_{\boldsymbol{\alpha}} \boldsymbol{L}(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{\alpha})$

$$\boldsymbol{if}\ \mathbf{1} > \boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + \boldsymbol{b})] \rightarrow \textbf{min won't let it happen}$$

$$\boldsymbol{if}\ \mathbf{1} = \boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + \boldsymbol{b})] \rightarrow \textbf{equivalent to} \min_{\boldsymbol{\theta},\boldsymbol{b}} \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta}$$

$$\boldsymbol{if}\ \mathbf{1} < \boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + \boldsymbol{b})] \rightarrow \textbf{equivalent to} \min_{\boldsymbol{\theta},\boldsymbol{b}} \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta}$$

LAGRANGIAN DUALITY

Primal

$$\min_{\boldsymbol{\theta},\boldsymbol{b}} \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta}$$

$$\text{s.t.} \quad \boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}+\boldsymbol{b}) \geq \mathbf{1}, \forall \boldsymbol{i}$$

The Lagrangian for the primal problem:

$\boldsymbol{L} = \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} + \sum_i \boldsymbol{\alpha}_i[\mathbf{1} - \boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}+\boldsymbol{b})]$ **where** $\boldsymbol{\alpha}_i \geq \mathbf{0}$ are Lagrange multipliers.

Now we want to solve: $\min_{\boldsymbol{\theta},\boldsymbol{b}} \max_{\boldsymbol{\alpha}} \boldsymbol{L}(\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{\alpha})$

Slater's condition from convex optimization guarantees that these two optimization problems are equivalent!

This is the dual problem $\max_{\boldsymbol{\alpha}} \min_{\boldsymbol{\theta},\boldsymbol{b}} \boldsymbol{L}(\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{\alpha})$

LAGRANGIAN DUALITY

Primal

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \boldsymbol{\theta}^T \boldsymbol{\theta}$$

$$\text{s.t.} \quad \boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + \boldsymbol{b}) \geq \mathbf{1}, \forall \boldsymbol{i}$$

$$\max_{\boldsymbol{\alpha}} \min_{\boldsymbol{\theta}, \boldsymbol{b}} \frac{1}{2} \boldsymbol{\theta}^T \boldsymbol{\theta} + \sum_{\boldsymbol{i}} \boldsymbol{\alpha}_{\boldsymbol{i}} [\mathbf{1} - \boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + \boldsymbol{b})]$$

Dual

$$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{\theta}} = \boldsymbol{\theta} - \sum_{\boldsymbol{i}} \boldsymbol{\alpha}_{\boldsymbol{i}} \boldsymbol{y}^{(i)} \boldsymbol{x}^{(i)} = \mathbf{0} \qquad \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{b}} = -\sum_{\boldsymbol{i}} \boldsymbol{\alpha}_{\boldsymbol{i}} \boldsymbol{y}^{(i)} = \mathbf{0}$$

LAGRANGIAN DUALITY

Primal

$$\min_{\boldsymbol{\theta}} \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta}$$

$$\text{s.t.} \quad \boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + \boldsymbol{b}) \geq \mathbf{1}, \forall \boldsymbol{i}$$

$$\max_{\boldsymbol{\alpha}} \boldsymbol{L}(\boldsymbol{\alpha}) = \frac{1}{2}\sum_{i}\sum_{j}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j\boldsymbol{y}^{(i)}\boldsymbol{y}^{(j)}\boldsymbol{x}^{(i)^T}\boldsymbol{x}^{(j)} + \sum_{i}\boldsymbol{\alpha}_i - \sum_{i}\sum_{j}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j\boldsymbol{y}^{(i)}\boldsymbol{y}^{(j)}\boldsymbol{x}^{(i)^T}\boldsymbol{x}^{(j)}$$

$$= \sum_i \boldsymbol{\alpha}_i - \frac{1}{2}\sum_i\sum_j \boldsymbol{\alpha}_i\boldsymbol{\alpha}_j\boldsymbol{y}^{(i)}\boldsymbol{y}^{(j)}\boldsymbol{x}^{(i)^T}\boldsymbol{x}^{(j)}$$

Dual

$$\text{s.t.} \quad \boldsymbol{\alpha}_i \geq \mathbf{0}, \forall \boldsymbol{i}$$

$$\sum_i \boldsymbol{\alpha}_i\boldsymbol{y}^{(i)} = \mathbf{0}$$

LAGRANGIAN DUALITY (SOFT MARGIN)

Primal

$$\min_{\theta,\xi,b} \frac{1}{2}\theta^T\theta + C\sum_{i=1}^{N}\xi_i$$

$$\text{s.t.}\quad y^{(i)}(\theta^T x^{(i)} + b) + \xi_i \geq 1,$$

$$\xi_i \geq 0, \forall i$$

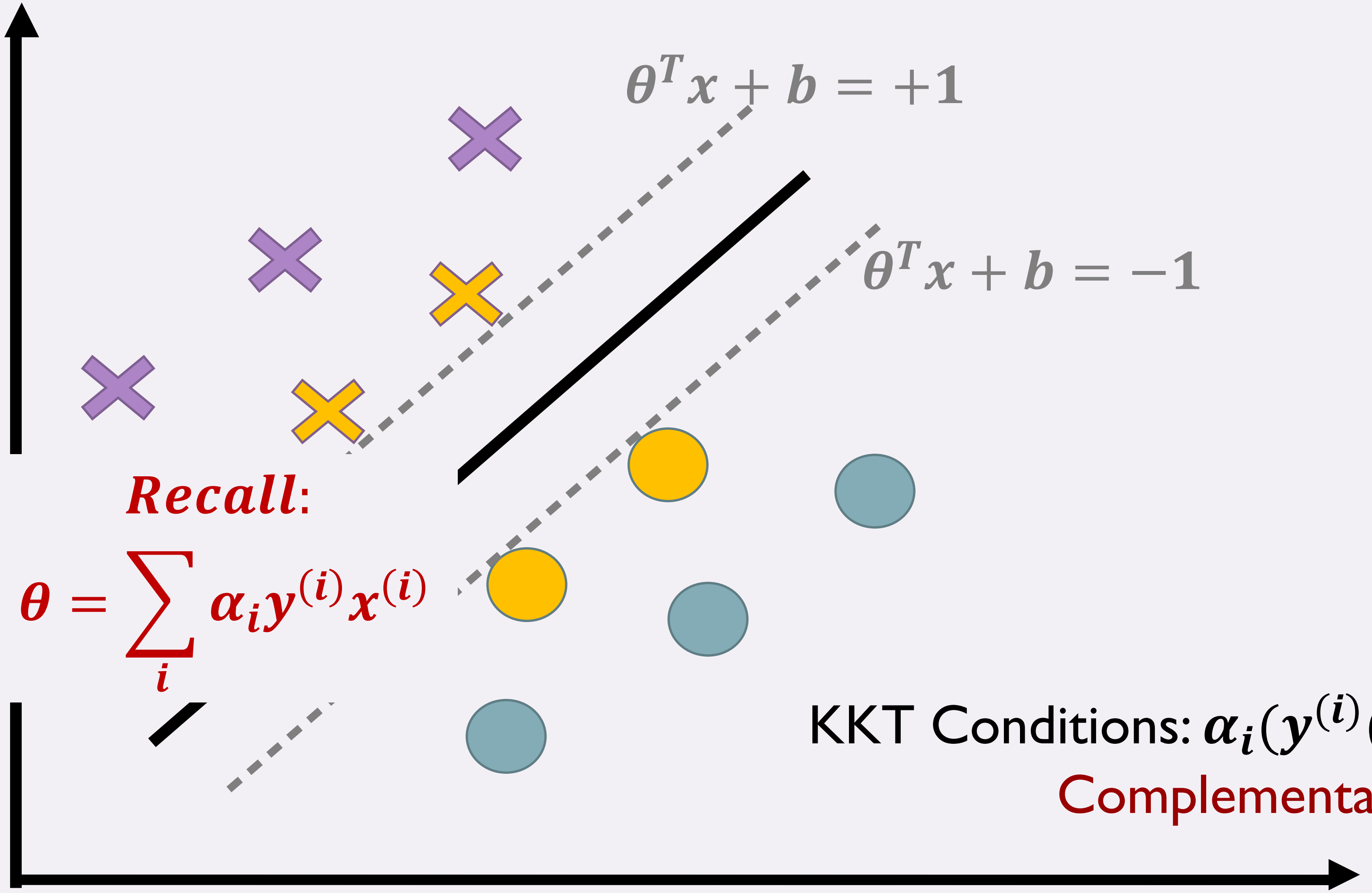
Dual

$$\max_{\alpha} L(\alpha) = \sum_i \alpha_i - \frac{1}{2}\sum_i\sum_j \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)^T} x^{(j)}$$

$$\text{s.t.}\quad \boxed{C} \geq \alpha_i \geq 0, \forall i$$

$$\sum_i \alpha_i y^{(i)} = 0$$

SUPPORT VECTORS



Recall:

$$\boldsymbol{\theta} = \sum_i \alpha_i y^{(i)} x^{(i)}$$

KKT Conditions: $\alpha_i(y^{(i)}(\theta^T x^{(i)} + b) - 1) = 0$

Complementary Slackness

KERNEL TRICK

$$\mathrm{x} \rightarrow \phi(x) \qquad R^d \rightarrow R^D$$

$$\min_{\boldsymbol{\theta}} \frac{\mathbf{1}}{\mathbf{2}}\, \boldsymbol{\theta}^{\boldsymbol{T}} \boldsymbol{\theta}$$

s.t. $\boldsymbol{y}^{(\boldsymbol{i})}(\boldsymbol{\theta}^{\boldsymbol{T}} \boldsymbol{\phi}(\boldsymbol{x}^{(\boldsymbol{i})}) + \boldsymbol{b}) \geq \mathbf{1}, \forall \boldsymbol{i}$

If D >> d then there are many more parameters to learn for θ Can this be avoided?

KERNEL TRICK

$$\mathrm{x} \rightarrow \phi(x) \qquad R^d \rightarrow R^D$$

$$\max_{\boldsymbol{\alpha}} \boldsymbol{L}(\boldsymbol{\alpha}) = \sum_{\boldsymbol{i}} \boldsymbol{\alpha_i} - \frac{\mathbf{1}}{\mathbf{2}} \sum_{\boldsymbol{i}} \sum_{\boldsymbol{j}} \boldsymbol{\alpha_i \, \alpha_j y^{(i)} y^{(j)}} \boldsymbol{\phi}(\boldsymbol{x}^{(i)})^{\boldsymbol{T}} \boldsymbol{\phi}(\boldsymbol{x}^{(j)})$$

s.t. $\boldsymbol{\alpha_i} \geq \mathbf{0}, \forall \boldsymbol{i}$

$\sum_{\boldsymbol{i}} \boldsymbol{\alpha_i y^{(i)}} = \mathbf{0}$

- $\phi(x)$ only occurs in pairs $\phi(x)^T \, \phi(x)$
- Once the scalar products are computed, only the N dimensional vector α needs to be learnt; it is not necessary to learn in the D dimensional space, as it is for the primal

KERNEL TRICK

Write $k\left(x_j, x_i\right) = \phi(x)^T\ \phi(x)$. This is known as **kernel**

$$\max_{\boldsymbol{\alpha}} \boldsymbol{L}(\boldsymbol{\alpha}) = \sum_{\boldsymbol{i}} \boldsymbol{\alpha_i} - \frac{\mathbf{1}}{\mathbf{2}} \sum_{\boldsymbol{i}} \sum_{\boldsymbol{j}} \boldsymbol{\alpha_i}\, \boldsymbol{\alpha_j} \boldsymbol{y}^{(\boldsymbol{i})} \boldsymbol{y}^{(\boldsymbol{j})} \boldsymbol{k}(\boldsymbol{x}^{(\boldsymbol{i})}, \boldsymbol{x}^{(\boldsymbol{j})})$$

s.t. $\boldsymbol{\alpha_i} \geq \mathbf{0}, \forall \boldsymbol{i}$

$\sum_{\boldsymbol{i}} \boldsymbol{\alpha_i} \boldsymbol{y}^{(\boldsymbol{i})} = \mathbf{0}$

KERNEL TRICK

$$\Phi : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \to \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{pmatrix} \quad \mathbb{R}^2 \to \mathbb{R}^3$$

- Classifier can be learnt and applied without explicitly computing Φ (x)
- All that is required is the kernel $k(x, z) = (x^T z)^2$

EXAMPLE KERNELS

- Linear $k(x, z) = x^T z$
- Polynomial of degree exactly d $k(x, z) = (x^T z)^d$
- Polynomial of degree up to d $k(x, z) = (x^T z + 1)^d$
- Guassian $k(x, z) = \exp(-||x-z||^2)$

  --- Infinite dimensional feature space

PROPERTIES

- Weight vector is a linear combination of data
- Only the points on the margins matter
  - Ignore the rest
- Only inner products between data points matter
  - Can use Kernels!
- Provides the widest possible separation between classes

QUIZ 3 (DUE TONIGHT)

Quiz 3

Not available until May 4 at 3:00pm | 4 pts | 5 Questions

HW2 (DUE MAY 9)

HW2

Available until May 9 at 11:59pm | **Due** May 9 at 11:59pm | 60 pts

QUESTIONS?