Make up class : Friday 31^{st}
Select time on Canvas

Learning with kernels, SVMs with Gaussian kernels on MNIST data (USPS)

homework 1 (Sunday)

Goal
- Nonlinearly separable datasets
- how to extend SVMs & kernel SVMs

Vanilla SVM

$$\min_{\substack{w \in \mathbb{R}^d,\ b \in \mathbb{R}}} \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad y_i(w\cdot x_i + b) \geq 1 \quad \forall i \in [m]$$

m: # of data pts

d : $\dim(\mathcal{X})$

Soln: $w = \sum_{i=1}^{m} \lambda_i y_i x_i$

$\lambda_i \neq 0$, x_i : support vectors

$\lambda = [\lambda_1, \ldots, \lambda_m] \in \mathbb{R}^m$ dual variables

Lagrangian:

$$\mathcal{L}(w,b,\lambda) = \frac{\|w\|^2}{2} + \sum_{i=1}^{m} \lambda_i (1 - y_i(w\cdot x_i + b))$$

Dual problem : obtained by plugging in KKT conditions

$$w = \sum_{i=1}^{m} \lambda_i y_i x_i$$

$$\sum_{i=1}^{m} y_i \lambda_i = 0$$

$\lambda_i = 0$ or $y_i(w\cdot x_i + b) = 1$

Dual problem:

$$\max_{\lambda} \sum_{i=1}^{m} \lambda_i - \frac{1}{2}\sum_{i=1}^{m} \lambda_i \lambda_j y_i y_j\, x_i \cdot x_j$$

$\lambda_i \geq 0 \quad i \in [m]$

In dot product form;
Replace $x_i \cdot x_j \rightarrow \Phi(x_i)\cdot\Phi(x_j)$

kernel SVM:

$$\max_{\lambda} \sum_{i=1}^{m} \lambda_i - \frac{1}{2}\sum_{i=1}^{m} \lambda_i \lambda_j y_i y_j\, \mathcal{K}(x_i, x_j)$$

$$h_{\mathcal{K}\text{-SVM}}(x) = \operatorname{sgn}\Big(\sum_{i=1}^{m} \lambda_i y_i \mathcal{K}(x_i, x) + b\Big)$$

To solve for b :
for any i at which $\lambda_i \neq 0$,
from KKT conditions,

$$y_i(w\cdot\Phi(x_i) + b) = 1$$

$$w\cdot\Phi(x_i) + b = y_i$$

$$b = y_i - w\cdot\Phi(x_i)$$

$$= y_i - \Phi(x)\cdot\sum_{i=1}^{m} \lambda_i y_i \Phi(x_i)$$

$$b = y_i - \sum_{i=1}^{m} \lambda_i y_i \mathcal{K}(x_i, x)$$

Kernelization of the SVM

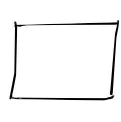
$x \rightarrow \Phi(x)$
data → features

dim d → $D \gg d$, ∞

classifier linear → linear model on feature space
equivalent ???
nonlinear model on data space

Recap: Cover's theorem:
in higher dim space, data are more likely to be linearly separable.

MNIST dat set



$28 \times 28 = 784$

$d = 784$

intrinsic dimension: $O(10)$

Linear classifiers not better than random guess!

Recap		hyperparameter

Gaussian kernel (PD)	$e^{-\frac{\|x-x'\|^2}{2\varepsilon^2}}$	ε
Polynomial kernel (PD)	$(x\cdot x' + c)^k$	c, k
Sigmoid kernel	$\tanh(a\, x\cdot x' + b)$	a, b

b, $a < 0$ not PD

PD kernel : • $\mathcal{K}$ is symmetric
• Gram matrix $G_{ij} = \mathcal{K}(x_i, x_j)$ is PSD. for all $m \in \mathbb{N}$.

Mercer's theorem
If
$\mathcal{K}(x, x') = \langle \Phi(x), \Phi(x') \rangle$ (Euclidean or ℓ^2 inner product)

then $\mathcal{K}$ is PD
Converse also holds.

Proof : if $\mathcal{K}$ is s.t. $\mathcal{K}(x,x') = \langle \Phi(x), \Phi(x') \rangle$

then,
$c^T G c \geq 0$ for any $c \in \mathbb{R}^m$.

$$\sum_{i,j=1}^{m} c_i c_j G_{ij} = \sum_{i,j=1}^{m} c_i c_j \Phi(x_i)\cdot\Phi(x_j)$$

$$= \Big\|\sum_{i=1}^{m} c_i \Phi(x_i)\Big\|^2 \geq 0.$$

Only if :
G is SPSD.

- $c^T G c \geq 0$ for any $c \in \mathbb{R}^m$.
- $G = P^T \Lambda P$ $\quad P^T = P^{-1}$, $\Lambda_i \geq 0$

$G_{ij} = \langle \sqrt{\Lambda_i} P_i, \sqrt{\Lambda_j} P_j \rangle$

$$\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} \Lambda_1 & \\ & \Lambda_2 \end{bmatrix} \begin{bmatrix} P_{11} & P_{21} \\ P_{12} & P_{22} \end{bmatrix}$$

$$\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} \Lambda_1 P_{11} & \Lambda_1 P_{21} \\ \Lambda_2 P_{12} & \Lambda_2 P_{22} \end{bmatrix} = \begin{bmatrix} \sqrt{\Lambda_1} P_{11}^2 \sqrt{\Lambda_1} + \sqrt{\Lambda_2}\sqrt{\Lambda_2} P_{12}^2 & \\ \bigcirc & \end{bmatrix}$$

$$P_{21}\Lambda_1 P_{11} + P_{22}\Lambda_2 P_{12}$$

$$\sqrt{\Lambda_1} P_{11} \cdot \sqrt{\Lambda_1} P_{21} + \sqrt{\Lambda_2}\sqrt{\Lambda_2} P_{22} P_{12}$$

$\Phi(x_i) = \sqrt{\Lambda_i}\, P_i \rightarrow$ i^{th} row of matrix P (whose columns are eigenvectors of G)

(Remember analog in ∞ dims)

$$\Phi(x_i)\cdot\Phi(x_j) = G_{ij}.$$

- Φ features are defined only at data points but they may not be "linearly" extensible to $\mathcal{X}$.

- Even if $\mathcal{K}$ is not PD but G is SPSD, we may still use features (extracted above) to "kernelize" algorithms in dot product form

---

Nonlinearly separable data

$\nexists$ w, b s.t. $y_i(w\cdot x_i + b) > 0$ at all datapts x_i, y_i

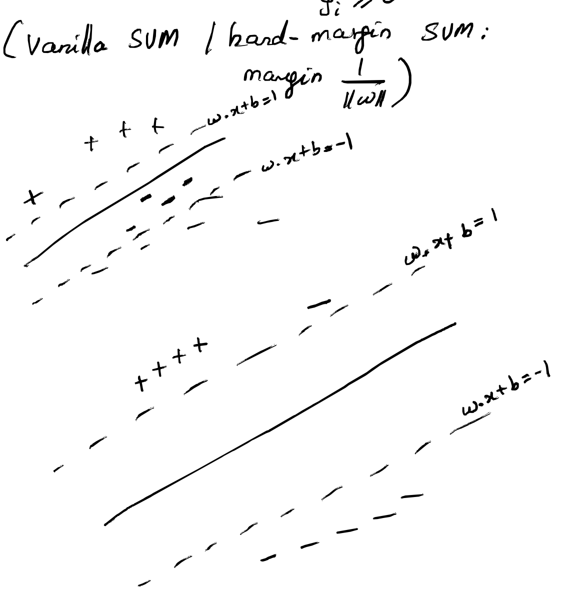
$$y_i(w\cdot\Phi(x_i) + b) > 0$$

Soft-margin SVM

$$\min_{w,b,\xi} \frac{\|w\|^2}{2} + C\|\xi\|^p$$

s.t. $y_i(w\cdot x_i + b) \geq 1 - \xi_i$
$\xi_i \geq 0$

(Vanilla SVM / hard-margin SVM: margin $\frac{1}{\|w\|}$)



$$\min_{w,\xi,b} \frac{\|w\|^2}{2} + C\|\xi\|^p \; : \text{obj. of soft margin classification}$$

- if C is large, we penalize "outliers"
- if C is small, we focus on max. margin

$p = 1$ or $p = 2$

$$\min_{w,\xi,b} \frac{\|w\|^2}{2} + C\sum_{i=1}^{m}\xi_i$$

s.t. $y_i(w\cdot x_i + b) \geq 1 - \xi_i$
$\xi_i \geq 0$

$$\mathcal{L}(w, b, \xi, \lambda, \gamma) = \frac{\|w\|^2}{2} + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\lambda_i\big(y_i(w\cdot x_i + b) - 1 + \xi_i\big) - \sum_{i=1}^{m}\gamma_i\xi_i$$

KKT

$$w = \sum_{i=1}^{m}\lambda_i y_i x_i$$

$$\sum_{i=1}^{m}\lambda_i y_i = 0$$

$$C = \lambda_i + \gamma_i$$

$y_i(w\cdot x_i + b) = 1 - \xi_i$ or $\lambda_i = 0$

$\gamma_i = 0$ or $\xi_i = 0$

Ex: kernel soft margin SVM dual problem

Support vectors x_i for which $\lambda_i \neq 0$.
$y_i(x\cdot w + b) = 1 - \xi_i$

Two types of support vectors:

	Location of x_i
$\xi_i = 0$	marginal hyperplane
$\xi_i \neq 0$, $\lambda_i = C$	outlier