SVM

Support Vector Machines

- Parametric classifier
- Motivation: Geometric (maximizing the margin) Not probabalistic

Outline of our study

1) Two-class SVM
   - linearly separable
   - linearly nonseparable
   - "kernel Trick"
2) Multi-class SVMs
   - aggregate / combined algorithms
     - one vs All (OVA)
     - one vs one (OVO)
   - uncombined algorithms (not covered)

Two-class Linearly separable SVM

- Label space $Y = \{-1, +1\}$
- Feature vectors $X = \mathbb{R}^d$
- Training set $D = \{(x_1, y_1), \ldots (x_n, y_n)\}$
- Family of classifiers: "linear" classifiers

$$H = \{h(x) = \text{sign}(w^T x + b) \quad w \in \mathbb{R}^d, \; b \in \mathbb{R}\}$$

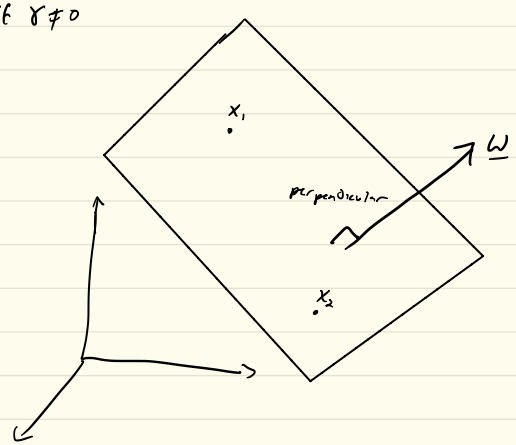
- Parameters

$$\theta = \begin{pmatrix} w \\ b \end{pmatrix} \quad (d+1) \text{ scalar parameters}$$

Equation of hyperplane in $\mathbb{R}^d$: $w^T x + b = 0$

(w, b) and $(\gamma w, \gamma b)$ define the same hyperplane if $\gamma \neq 0$

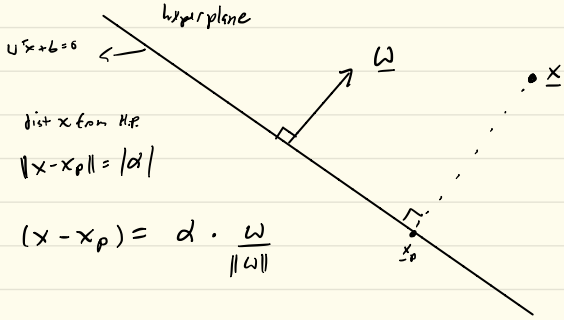
$(\gamma w)^T x + \gamma b = 0$, $\gamma \neq 0$



$w^T x_1 + b = 0$

$w^T x_2 + b = 0$

$w^T (x_1 - x_2) = 0$



$(x - x_p) = d \cdot \frac{w}{\|w\|}$

⊥ dist. (euclid.) of pt. x from hyperplane

$= \frac{|w^T x + b|}{\|w\|}$ ← euclidean

$w^T x_p + b = 0$

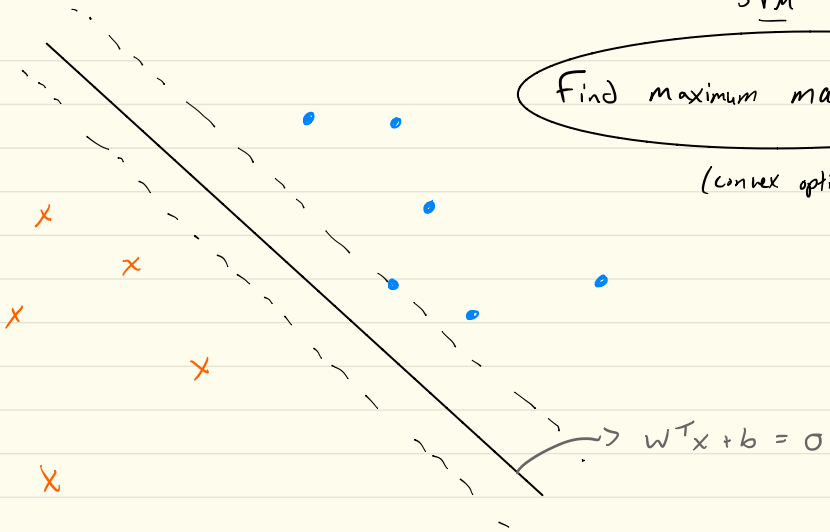
$\Rightarrow w^T (x - x_p) = d \frac{w^T w}{\|w\|} = d \frac{\|w\|^2}{\|w\|} = d\|w\|$

$\Rightarrow d = \frac{w^T w - w^T x_p}{\|w\|} = \frac{w^T x + b}{\|w\|} \Rightarrow |d| = \frac{|w^T x + b|}{\|w\|}$

SVM

Find maximum margin hyperplane

(convex optimization problem)



- Geometric margin of a hyperplane (w, b) wrt D

$$\rho(w,b) := \min_{1 \le j \le n} \frac{|w^T x_j + b|}{\|w\|}$$

(distance of feature vector which is closest to the hyperplane)

- Linearly separable training set D:

$$\exists_{\text{exists}} (w,b) \text{ s.t. } y_j (w^T x_j + b) > 0 \quad \forall j = 1, \dots, n$$

$$\Rightarrow \boxed{\rho(w,b) > 0}$$

Canonical Parameterization of hyperplane (w,b) wrt lin. sep. D:

If D is lin. sep. by (w,b) and $\rho > 0$, then the canonical parameterization (w_{canon}, b_{canon}) is such that

$$\min_{1 \le j \le n} \left| w_{canon}^T x_j + b_{canon} \right| = 1$$

for canonical parameterization : $\rho(w,b) = \frac{1}{\|w\|}$

$$(w_{svm}, b_{svm}) = \underset{(w,b):}{\operatorname{argmax}} \quad \rho(w,b)$$

1) Perfect lin. sep. of D
2) Canonical rep.

$$(w_{svm}, b_{svm}) = \underset{(w,b):}{\operatorname{argmax}} \quad \frac{1}{\|w\|}$$

1) $y_j(w^T x_j + b) > 0 \quad \forall_j$
2) $\min_{1 \le j \le n} |w^T x_j + b| = 1$

$$\rho_{opt} = \rho(w_{svm}, b_{svm}) = \max_{(w,b)} \frac{1}{\|w\|}$$

1) ...
2) ...

$$\rho_{opt} = \max_{(w,b):} \frac{1}{\|w\|} = \max_{(w,b):} \frac{1}{\|w\|}$$

1) $y_j(w^T x_j + b) > 0 \quad \forall_j$ $\qquad y_j(w^T x_j + b) \ge 1 \ \forall_j$
2) $\min_j |w^T x_j + b| = 1$

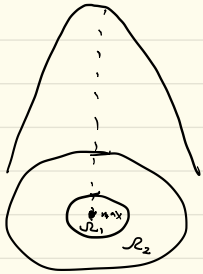
$2) \Rightarrow \quad |w^T x_j + b) \ge 1$

$= \quad y_j(w^T x_j + b) \ge 1$

(w_*, b_*) $\quad \frac{1}{\|w_*\|}$

$\min_{1 \le j \le l} |w_*^T x_j + b_*| > 1$

take $(w_{svm}, b_{svm}) = \frac{(w_*, b_*)}{\gamma}$



$\Rightarrow (w_{svm}, b_{svm})$ $\quad \underset{(w,b):}{\text{argmax}} \quad \frac{1}{\|w\|}$

$\forall j, \; y_j (w^T x_j + b) \ge 1$

$(w_{svm}, b_{svm}) = \underset{(w,b):}{\text{argmin}} \quad \frac{1}{2}\|w\|$

$\forall j \quad y_j (w^T x_j + b) \ge 1$

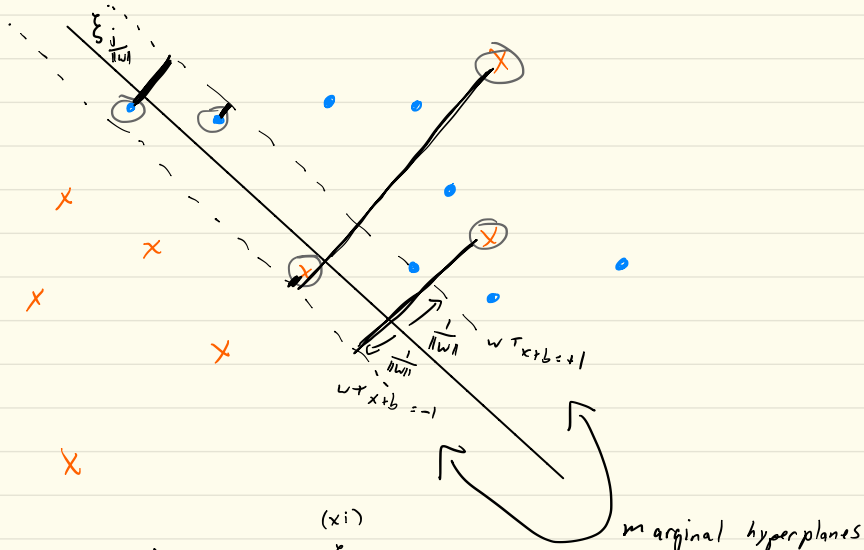
$= \underset{(w,b)}{\text{argmin}} \quad f(w,b)$

$\forall j \quad g_j(w,b) \ge 0$

$f(w,b) = \frac{1}{2}\|w\|^2$

$g_j(w,b) = 1 - y_j (w^T x_j + b)$

non-separable



$$y_j (w^T x_j + b) \geq 1 - \xi_j$$ (ξ_j: slack for j)

objective function of SVM:

$$(w_{svm}, b_{svm}) = \underset{(w,b)}{\operatorname{argmin}} \left[\frac{1}{2}\|w\|^2 + \text{penalty}(\xi_1, \ldots, \xi_n) \right]$$

$$\forall_j \quad y_j (w^T x_j + b) \geq 1 - \xi_j, \quad \xi_j \geq 0$$

Penalty Examples

1) $C\left(\sum_{j=1}^{n} \xi_j\right), \quad C > 0$

2) $C \sum_{j=1}^{n} (\xi_j)^p, \quad C > 0, \; p > 0$

3) $C_+ \sum_{j: y_j = +1} \xi_j + C_- \sum_{j: y_j = -1} \xi_j$

$$w_{svm} = \sum_{j=1}^{n} \alpha_j^* y_j x_j, \quad 0 \leq \alpha_j^* \leq C_j \; (\forall_j)$$

$$b_{svm} = y_j - w_{svm}^T x_j \text{ for any } j : 0 < \alpha_j^* < C_j$$

Typically $n_+ C_+ = n_- C_- = C$

$\Rightarrow C_+ = \frac{C}{n_+}, \; C_- = \frac{C}{n_-}$

n_+ = # pos samples, n_- = # neg samples

General case:

$$\text{Penalty} = \sum_{j=1}^{n} c_j \xi_j, \quad c_j > 0 \ \forall j$$

Primal optimization problem

$$(w_{svm}, b_{svm}) = \underset{(w,b)}{\operatorname{argmin}} \left[\frac{1}{2}\|w\|^2 + \sum_{j=1}^{n} c_j \xi_j \right]$$

$$\forall j \left[\begin{array}{c} y_j (w^T x_j + b) \geq 1 - \xi_j \\ \xi_j \geq 0 \end{array} \right]$$

$\xi_j \rightarrow$ "slack" variable

$c_j > 0 \rightarrow$ trade off between maximizing soft margin $1/\|w\|$ and minimizing degree of ideal constraint violation (slack)

Solution

- convex optimization problem: quadratic program (QP)
- unique global minimizer given by:

$$w_{svm} = \sum_{j=1}^{n} \alpha_j^* y_j x_j$$

Satisfying the following Complementary Slackness conditions

1) $$\forall j \quad \alpha_j^* \left[\begin{array}{c} y_j (w_{svm}^T x_j + b_{svm}) \\ -1 + \xi_j^* \end{array} \right] = 0$$

2) $$(c_j - \alpha_j^*) \xi_j^* = 0 \quad \forall j$$

$$b_{svm} = y_j - w_{svm}^T x_j \qquad \text{for any } j: 0 < \alpha_j^* < c_j$$

& $\underline{\alpha}^* = (\alpha_1^*, \ldots, \alpha_n^*)^T$

is the solution to the following DUAL optimization problem

Solution

- convex optimization problem: quadratic program (QP)
- unique global minimizer given by:

$$w_{svm} = \sum_{j=1}^{n} \alpha_j^* y_j x_j$$

satisfying the following Complementary Slackness conditions

1) $\forall_j \quad \alpha_j^* \left[y_j (w_{svm}^T x_j + b_{svm}) - 1 + \xi_j^* \right] = 0$

2) $(C_j - \alpha_j^*) \xi_j^* = 0 \quad \forall_j$

$b_{svm} = y_j - w_{svm}^T x_j$ for any $j: 0 < \alpha_j^* < C_j$

& $\underline{\alpha}^* = (\alpha_1^*, \ldots, \alpha_n^*)^T$

is the solution to the following DUAL optimization problem

$$\underline{\alpha}^* = \arg\max_{\underline{\alpha}:} \left[\sum_{j=1}^{n} \alpha_j - \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} \alpha_j \alpha_k y_j y_k (x_j^T x_k) \right]$$

$\forall_j, \quad 0 \le \alpha_j \le C_j$

$\sum_{j=1}^{n} \alpha_j y_j = 0$

can be solved for using sequential minimal optimization (SMO) algorithm

$O(n^\gamma) \quad \gamma \in [2, 3]$

$\{ x_j : \alpha_j > 0 \} \rightarrow$ Support Vectors

$\{ x_j : \xi_j > 0 \} \rightarrow$ Outliers

All outliers are support vectors
→ from complementary slackness condition #2 : $\xi_j^* > 0 \Rightarrow \alpha_j^* = C_j > 0$

Support vectors $\Rightarrow \alpha_j^* \neq 0$

then by Comp. slackness cond #1

$$y_j (w_{SVM}^T x_j + b_{SVM}) - 1 + \xi_j^* = 0$$

If $\xi_j^* = 0$
(not outlier)
$y_j (w_{SVM}^T x_j + b_{SVM}) = +1$

$\xi_i^* > 0$
$\Rightarrow$ outlier

Unconstrained Form of the SVM Primal optimization problem

$$\forall j \left[y_j (w^T x_j + b) \geq 1 - \xi_j , \ \xi_j \geq 0 \right]$$

$$\forall j \left[\xi_j \geq 1 - y_j (w^T x_j + b) , \ \xi_j \geq 0 \right]$$

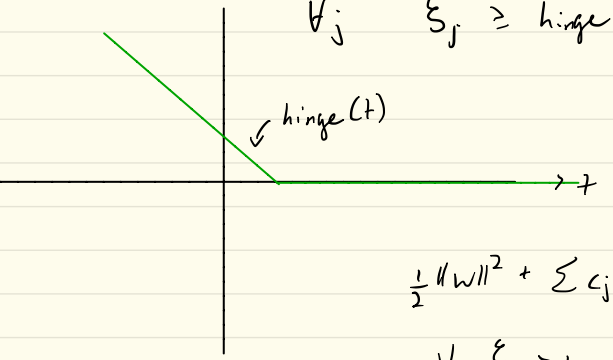
$$= \forall j \left[\xi_j \geq \max (0, \ 1 - y_j (w^T x_j + b)) \right]$$

$\text{hinge}(y_j(w^T x_j + b))$

$\text{hinge}(t) = \max(0, 1-t)$

$(w_{svm}, b_{svm}) = \underset{(w,b)}{\text{argmin}} \left[\frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} c_j \xi_j \right]$

$\forall j \quad \xi_j \geq \text{hinge}(y_j(w^T x_j + b))$



$\frac{1}{2}\|w\|^2 + \sum c_j \xi_j \quad > \quad \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} c_j \text{hinge}(y_j(w^T x_j + b))$

$\forall j, \ \xi_j \geq \text{hinge}(y_j(w^T x_j + b))$

ex.

$\text{cost} = \frac{1}{2}\|w\|^2 + c_1 \text{hinge}(1) + c_1 \text{hinge}(1)$

$+ c_2 \text{hinge}(2)$

$c_1 \to 2c_1$

w_{svm} is <u>unique</u>

proof:

$$\begin{pmatrix} w \\ b \end{pmatrix} \in \min_{C} \left[f(w) + g(b) \right]$$

- C: convex set
- $f(w)$: STRICTLY convex
- $g(b)$: convex

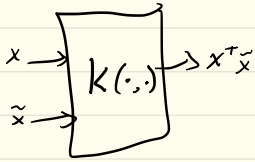
b_{svm} is unique if D is linearly separable

If D is not lin. sep. there may be multiple (even infinitely many) solutions for b_{svm}

$$w_{svm} = \begin{pmatrix} C \\ 0 \end{pmatrix}$$

$$h_{svm}(x) = \text{sign}\left(w_{svm}^T x + b_{svm}\right)$$

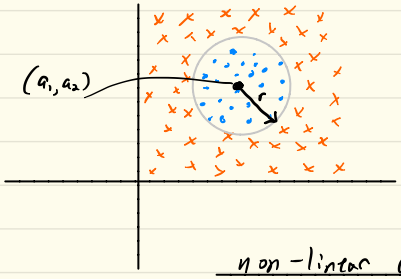
$$= \text{sign}\left(\sum_{i=1}^{n} \alpha_i^* y_i (x_i^T x) + b_{svm}\right)$$



Kernel Trick Motivation

- D may not be linearly separable, but may be so "non linearly"

Example



$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

non-linear classifier

$$(x_1 - a_1)^2 + (x_2 - a_2)^2 \underset{-1}{\overset{+1}{\lessgtr}} r^2$$

$$a_1^2 + x_2^2 - 2a_1x_1 - 2a_2x_2 + \underbrace{(a_1^2 + a_2^2 - r^2)}_{b} \underset{-1}{\overset{+1}{\lessgtr}} 0$$

$$\underbrace{\begin{pmatrix} 1 & 1 & -2a_1 & -2a_2 \end{pmatrix}}_{w^T} \underbrace{\begin{pmatrix} x_1^2 \\ x_2^2 \\ x_1 \\ x_2 \end{pmatrix}}_{\phi\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right)}$$

$$\underbrace{w^T \phi(x)}_{\in \mathbb{R}^4} + b \underset{-1}{\overset{+1}{\lessgtr}} 0$$

$$D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

$$\downarrow \phi(\cdot)$$

$$D_{new} = \{(\phi(x_1), y_1), \ldots, (\phi(x_n), y_n)\}$$

Idea replace $x \in \mathbb{R}^d$ with $\phi(x) \in \mathbb{R}^{d'}$, $d' > d$, to make D more linearly separable

→ But this also increases computational complexity (dot prods now in $\mathbb{R}^{d'}$)

→ We also need to construct $\phi(\cdot)$

Solution: Positive Definite Symmetric (PDS) kernels

→ these IMPLICITLY define an inner product in an IMPLICIT high-dim space (could even be ∞-dimensional)

PDS kernel definition

$k(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is PDS if

1) Symmetric: $\forall\, x, \tilde{x} \in \mathbb{R}^d$
$k(x, \tilde{x}) = k(\tilde{x}, x)$

2) $\forall n$ & all $x_1, \ldots, x_n \in \mathbb{R}^d$
the $n \times n$ matrix

$$K_n = \begin{bmatrix} k(x_1,x_1) & \cdots & k(x_1,x_n) \\ \vdots & \ddots & \vdots \\ k(x_n,x_1) & \cdots & k(x_n,x_n) \end{bmatrix}$$

is positive semi-definite

i.e. $\forall\, c \in \mathbb{R}^n$, $c^T K_n c \geq 0$

Result (Mercer's theorem in functional analysis)

If $K(\cdot,\cdot)$ is PDS then

$\exists \quad \phi : \mathbb{R}^d \to \underbrace{V}$

Hilbert space

V: vector space of suitable dimension (could be ∞)

Such that $\forall\, x, \tilde{x} \in \mathbb{R}^d$

$$K(x, \tilde{x}) = \langle \phi(x), \phi(\tilde{x}) \rangle$$

($\langle \cdot, \cdot \rangle$: suitable inner product in V)

Examples of PDS kernels

1) **Polynomial**

$$K(x, \tilde{x}) := (x^T \tilde{x} + c)^\ell$$

$c > 0$ offset, ℓ = positive integer

Special case $c = 0, \ell = 1$

$$K(x, \tilde{x}) = x^T \tilde{x}$$
$$= \langle \phi(x), \phi(\tilde{x}) \rangle$$

$\boxed{\phi(x) = x}$ Identity Transformation

- Gaussian Kernel / Radial Basis Function (RBF) kernel

$$k(x, \tilde{x}) := e^{\frac{-\|x-\tilde{x}\|^2}{2\sigma^2}}$$

$\sigma > 0$

- Normalized RBF

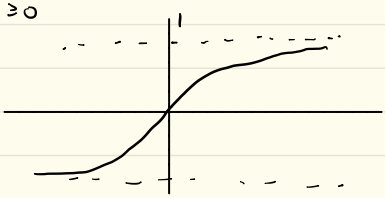
$$k(x, \tilde{x}) := e^{\frac{x^T \tilde{x}}{\sigma^2}}, \quad \sigma > 0$$

- Sigmoid kernel

$$k(x, \tilde{x}) := \tanh\left(a(x^T \tilde{x}) + b\right)$$

$a > 0, \quad b \geq 0$

$$\tanh(t) = \left(\frac{e^{+t} - e^{-t}}{e^{+t} + e^{-t}}\right)$$



Properties (PDS kernel algebra):

Let $k_1(\cdot,\cdot)$ & $k_2(\cdot,\cdot)$ be two PDS kernels

then the following new kernels derived from k_1 & k_2 are also PDS

1) Sum: $k_{sum}(x, \tilde{x}) := k_1(x, \tilde{x}) + k_2(x, \tilde{x})$

2) Positive scaling: $\alpha k_1(x, \tilde{x}), \quad \forall \alpha > 0$

3) Product: $k_{prod}(x, \tilde{x}) := \left(k_1(x, \tilde{x})\right)\left(k_2(x, \tilde{x})\right)$

4) exponentiation : $k_{exp}(x, \tilde{x}) := e^{k_1(x, \tilde{x})}$

5) Power-series : $k(x, \tilde{x}) = \sum_{\ell=0}^{\infty} a_\ell \left(k_1(x, \tilde{x})\right)^\ell$

($a_\ell \geq 0 \; \forall \ell$)

Ans

Multi-class SVMs

↳ Aggregate / combined algorithms

- ↳ One-vs.-all (OVA)
- ↳ One-vs.-one (OVO)
- ↳ Error-correction-classification (ECC)

One-vs-all (OVA)

Idea: m classes : $1, 2, \ldots, m$

Learn m Binary classifiers

For $k = 1, \ldots, m$

$$h_k(x) = \begin{cases} +1 \Rightarrow x \text{ classified as class } k \\ -1 \Rightarrow x \text{ classified as } \underline{\text{NOT}} \ k \end{cases}$$

For k^{th} classifier: relabel training data as follows

$$y_j(k) = \begin{cases} +1 & \text{if } y_j = k \\ -1 & \text{otherwise} \end{cases}$$

Final Decision

Typically $h_k(x) = \text{sign}(\underbrace{f_k(x)}_{\text{some scoring function}})$

$$h_{OVA}(x) = \underset{1 \le k \le m}{\text{argmax}} \ f_k(x)$$

issues 1) Unbalanced data sets

→ If all classes have approx. equal examples then $\forall k$, the # positive samples for k^{th} classifier $= \frac{n}{m}$

vs. $\frac{(m-1)n}{m}$ for negative

2) Uncalibrated score functions $f_1(x) \ldots f_m(x)$

→ not a probability (confidence)

So range of f_1 may be different from f_2 etc.

One-vs.-one (OVO)

m classes $\Rightarrow \binom{m}{2} = \frac{m(m-1)}{2}$ pairs of classes

- Learn $\binom{m}{2}$ binary classifiers.

$h_{\{k,\tilde{k}\}}(x)$ one for every unordered pair $\{k, \tilde{k}\}$, $k \neq \tilde{k}$

$$h_{\{k,\tilde{k}\}}(x) = \begin{cases} k & \text{if classified as } k \\ \tilde{k} & \text{otherwise} \end{cases}$$

$$h_{ovo}(x) = \underset{1 \le k \le m}{\arg\max} \underbrace{\left| \left\{ \tilde{k} \neq k : h_{\{k,\tilde{k}\}}(x) = k \right\} \right|}_{\text{\# of classes against which class } k \text{ 'wins'}}$$

↑ class with the largest number of wins

(+) → No calibration problem
Combining <u>decisions</u> not uncalib. scores

(+/-?) → Computational Complexity
- training
- testing

(-) → Fewer total # points for training any individual class $\left(\frac{2n}{m}\right)$

(+) → Datasets are balanced for each classifier

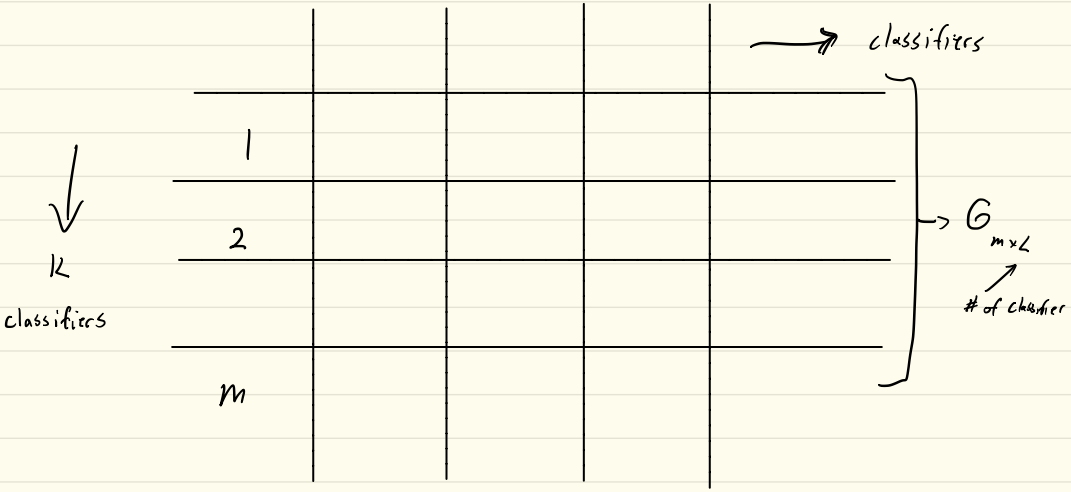
Computational Cost (rough) analysis

Assume $\approx \frac{n}{m}$ pts. in each class

$O(n^\gamma)$ complexity for training a single binary classifier in the family on a dataset of size n

C_{test} = Cost of testing with 1 binary

	training	testing
OVA	$O(m^1 n^\gamma)$	$O(mC_{test})$
OVO	$O(m^2(\frac{n}{m})^\gamma)$	$O(m^2 C_{test})$



Multi-class classification (not specific to SVM)

Combined / Aggregated Methods

- Family of binary classifiers
- trained on different labelings of training set
- scores / decisions are combined / aggregated to arrive at final multiclass label

- OVA
- OVO
- ECC

Uncombined multi-class SVM

2 class

$$(w_{svm}, b_{svm}) = \underset{(w,b,\xi)}{\operatorname{argmin}} \left[\frac{1}{2}\|w\|^2 + \sum_{j=1}^{n} c_j \xi_j \right]$$

$$\forall j : y_j (w^T x_j + b) \geq 1 - \xi_j$$

$$\xi \geq 0$$

$$h_{svm}(x) = \text{sign}\,(w_{svm}^T x + b_{svm})$$

$$h_{\text{multiclass svm}}(x) = \underset{1 \leq k \leq m}{\operatorname{argmax}} \left(w_{k,svm}^T x + b_{k,svm} \right)$$

$$(\underline{w}_{1,svm}, \underline{w}_{2,svm}, \dots, \underline{w}_{m,svm})$$

$$(b_{1,svm}, b_{2,svm}, \dots, b_{m,svm})$$

$$= \text{argmin} \left[\frac{1}{2} \sum_{k=1}^{m} \|w_k\|^2 + \sum_{j=1}^{m} c_j \xi_j \right]$$

$$\forall j, \forall \ell \neq y_j, \xi_j \geq 0,$$

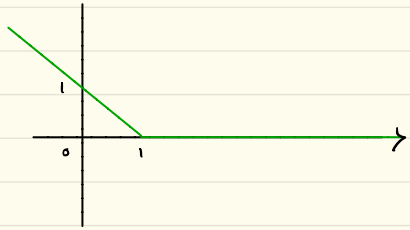
$$(w_{y_j}^T x + b_{y_j}) \geq (w_\ell^T x + b_\ell) + 1 - \xi_j$$

SVMs do not arise from either Generative / Discriminitive probabalistic models

Unconstrained Form of the SVM optimization problem

$$(w_{svm}, b_{svm}) = \underset{(w,b)}{\operatorname{argmin}} \left[\frac{1}{2} \|w\|^2 + \sum_{j=1}^{n} c_j \, \text{hinge}(y_j(w^T x_j + b)) \right]$$

$$\text{hinge}(t) = \max(0, 1-t)$$



Assume (only for simplicity) that $c_j = c > 0 \quad \forall j$

then minimizing becomes

$$\sum_{j=1}^{n} c \cdot \text{hinge}(y(w^T x_j + b)) \quad + \quad \frac{1}{2}\|w\|^2$$

Gen. : $-\ln p(x_i, y_i \mid \overbrace{w, b}^{\theta})$ $\qquad$ $-\ln \pi(w)$

Disc. : $-\ln p(y_i \mid x_i, w, b)$ $\qquad$ $\pi(w) \quad \frac{-\|w\|^2}{2}$

$= \frac{e}{(\sqrt{?})^?}$

Gen?

$$p(x, y \mid w, b) = \frac{1}{2?} e^{-c\,\text{hinge}(y(w^T x + b))}$$

Require

$$\sum_{y=\pm 1} \int_{x \in \mathbb{R}^2} p(x, y \mid w, b)\, dx = 1$$

$$\Rightarrow \frac{1}{2} \int_x \left[e^{-c\,\text{hinge}(+1(w^T x + b))} + e^{-c\,\text{hinge}(-(w^T x + b))} \right] dx = 1$$

$$Z =$$