LwP

1/12/19

μ^+: spam prototype

μ^-: non-spam prototype

x: feature vector of new email

x is classified as spam if

$$\|x-\mu^+\|_2 < \|x-\mu^-\|_2$$

$$\Rightarrow \|x-\mu^+\|_2^2 < \|x-\mu^-\|^2$$

$$\Rightarrow \|\mu^+\|_2^2 - 2\langle x, \mu^+\rangle < \|\mu^-\|_2^2 - 2\langle x, \mu^-\rangle$$

$$\Rightarrow \langle x, 2(\mu^+-\mu^-)\rangle + \|\mu^-\|_2^2 - \|\mu^+\|_2^2 > 0$$

$$\equiv \langle x, w\rangle + b > 0$$

with $w = 2(\mu^+-\mu^-)$

$$b = \|\mu^-\|_2^2 - \|\mu^+\|_2^2$$

→ decision boundary of LwP classifier is always a line or a hyperplane if the dist. fn. is Euclidean.

⇒ LwP is a linear classifier.

→ Model

- w
  - → vector of dim d, the same as that of features
  - → aka normal of the hyperplane
- b
  - → ka bias, it is a scalar term

→ Prediction: test point x by checking if the following holds

$$\boxed{w^Tx + b > 0}$$

→ Dec. boun: line/hyperplane where

$$\boxed{w^Tx + b = 0}$$

Pf: x, y 2 pt.s on hyperplane

$$\Rightarrow w^Tx + b = 0$$

$$\& \; w^Ty + b = 0$$

$$\Rightarrow w^T(x-y) = 0$$

acts as a threshold: how large does w^Tx have to be in order to ~~be~~ classify x as spam

→ Convenient to not have bias term

let $\tilde{x} = [x, 1] \in \mathbb{R}^{d+1}$

$\tilde{\omega} = [\omega_0, \omega_1, \cdots \omega_d] \in \mathbb{R}^{d+1}$

where $\omega = [\omega_0, \omega_1, \cdots \omega_{d-1}] \in \mathbb{R}^d$

then $\boxed{\tilde{\omega}^T \tilde{x} = \omega^T x + \omega_d}$ → effectively, acts as a bias term.

Metric learning : learning the distance fn.

↳ Eg. Mahalanobis metrics

let A: symm mat. $\in \mathbb{R}^{d \times d}$

def. $\boxed{d_A(x,y) = \sqrt{(x-y)^T A (x-y)}}$

→ A satisfies positive semi-definiteness (PSD)

$\forall x$, $\boxed{x^T A x \geq 0}$

$\Rightarrow d_A(x,y) \geq 0$

LwP w/ Mahalanobis

$(x-\mu^+)^T A (x-\mu^+) < (x-\mu^-)^T A (x-\mu^-)$

$d_A(x, \mu^+) < d_A(x, \mu^-)$

$\Rightarrow 2x^T A(\mu^+ - \mu^-) + \mu^{-T} A \mu^- - \mu^{+T} A \mu^+ > 0$

$\equiv \langle x, \omega \rangle + b > 0$

where $\omega = 2A(\mu^+ - \mu^-)$

$b = \mu^{-T} A \mu^- - \mu^{+T} A \mu^+$

Let $A = LL^T$, where L need not be sym. or PSD

Then $d_A(x,y) = \sqrt{(x-y)^T A (x-y)} = \sqrt{(x-y)^T L L^T (x-y)}$

$= \| L^T x - L^T y \|_2$

→ Mahalon. dist is the Euc. dist if x is transfm to Lx

Hyperparameter tuning

k in kNN
r in rNN
metric used in LwP
} hyperparameters of ML algo → tuned to give highest test accuracy.

w: model vector } parameters

hyp param: help in learning proc
param: used to make decisions (eg: wt. co-eff of lin.reg. model)

Held-out validation.

Multi-fold cross validation

→ finding NN for a new pt. is expensive
→ all training w/ NN needs to be stored, but only prototype stored in LwP
→ entire training set is the model for NN
→ model size in NN inc w/ the amt of training data; such models ka non-parametric models ✓
→ LwP model: just 2 vectors, indep of training set; such models ka parametric models ✓

Decision trees

→ faster way to perf. NN searches
→ model size can be large
→ good train perf. but bad train perf. (ka overfitting)
→ balanced DT's take $O(\log n)$ time for prediction

(L5)

Large Margin Classifiers

d of p from hyperplane $w^T x + b = 0$ is $\frac{|w^T p + b|}{\|w\|_2}$

Let train data be $\{(x^i, y^i)\}_{i=1}^n$

where $x^i \in \mathbb{R}^d$

$y^i \in \{-1, 1\}$

Classify each pt correctly : $\text{sign}(w^T x^i + b) = y^i$

or $y^i \cdot (w^T x^i + b) \geq 0$

Pt's away from boundary : $\min_{i=1\ldots n} \frac{|w^T x^i + b|}{\|w\|_2}$ shd be as large as possible

SVM : find a lin classif that
- → perf classifies data
- → keeps pt's as far away as possible

Objective $\max_{w,b} \left\{ \min_{i=1\ldots n} \frac{|w^T x^i + b|}{\|w\|_2} \right\}$ st $y^i \cdot (w^T x^i + b) \geq 0$ → Constraint

$\forall\, i = 1 \ldots n$

→ optimization prob w/ constraints.

$$\frac{|w^T x^i + b|}{\|w\|_2} = \frac{|y^i \cdot (w^T x^i + b)|}{\|w\|_2}$$

$$\min_{i=1\cdots n} \frac{|y^i \cdot (w^T x^i + b)|}{\|w\|_2} = \min_{i=1\cdots n} y^i \cdot \frac{(w^T x^i + b)}{\|w\|_2} \quad (\because \text{perf. classific})$$

let y^{i_0} be the closest pt

$$\Rightarrow \min_{i=1\cdots n} y^i \cdot (w^T x^i + b) = y^{i_0} \cdot (w^T x^{i_0} + b)$$

let $\varepsilon = y^{i_0} \cdot (w^T x^{i_0} + b)$

$\tilde{w} = w/\varepsilon$

$\tilde{b} = b/\varepsilon$

$$\Rightarrow y^i \cdot (\tilde{w}^T x^i + \tilde{b}) \geq 1 \quad \forall i = 1\cdots n$$

& $$\min_{i=1\cdots n} \frac{y^i \cdot (\tilde{w}^T x^i + \tilde{b})}{\|\tilde{w}\|_2} = \frac{1}{\|\tilde{w}\|_2}$$

thus $$\max_{\tilde{w},\tilde{b}} \left\{ \frac{1}{\|\tilde{w}\|_2} \right\} \text{ st } y^i \cdot (\tilde{w}^T x^i + \tilde{b}) \geq 1 \quad \forall i = 1\cdots n$$

$$\Downarrow$$

$$\min_{\tilde{w},\tilde{b}} \{ \|\tilde{w}\|_2 \}$$

$$\Downarrow$$

$$\min_{\tilde{w},\tilde{b}} \{ \|\tilde{w}\|_2^2 \}$$

$$\Downarrow$$

$$\min_{\tilde{w},\tilde{b}} \frac{1}{2} \{ \|\tilde{w}\|_2^2 \}$$

$$\text{st } y^i \cdot (w^T x^i + b) \geq 0$$

$$\left\{ \min \frac{|w^T x^i + b|}{\|w\|_2} \right\} \text{ st } y^i \cdot (w^T x^i + b) \geq 0$$

If the classific can't be perf^{n}, then model using

$$\min_{\tilde{w},\tilde{b},\{\xi_i\}} \frac{1}{2}\|\tilde{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i$$

st $y^i \cdot (\tilde{w}^T x^i + \tilde{b}) \geq 1 - \xi_i \quad \forall\ i = 1 \cdots n$

& $\xi_i \geq 0 \quad \forall\ i = 1 \cdots n$

↳ slack variables

Thus ξ_i allows $y^i \cdot (\tilde{w}^T x^i + \tilde{b}) \leq 1$

Thus we need $\xi_i = 1 - y^i \cdot (\tilde{w}^T x^i + \tilde{b})$

however $\xi_i \geq 0$

Thus if $y^i \cdot (\tilde{w}^T x^i + \tilde{b}) \geq 1$, $\xi_i = 0$ (no slack reqd)

Thus $\xi_i = \left[1 - y^i \cdot (\tilde{w}^T \cdot x^i + \tilde{b})\right]_+$

↳ $[x]_+ = \max\{x, 0\}$

Hinge loss fn

Hinge Loss - captures how well a data pt was classified

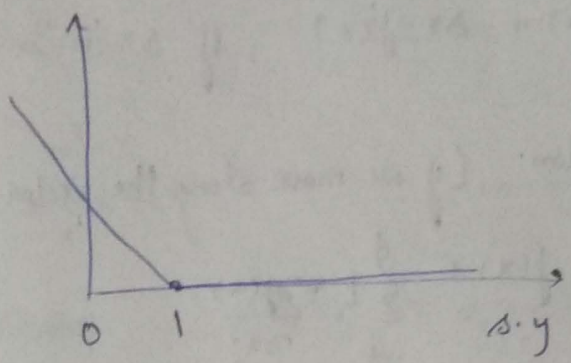
→ on data pt (x, y), $y \in \{-1, 1\}$

prediction score s (for lin model (w, b) we have, $s = w^T x + b$)

$s \cdot y \geq 0$ for correct classific

$s \cdot y >> 0$ for large margin

$$l_{hinge}(s,y) = [1 - s \cdot y]_+ = \begin{cases} 0 & \text{if } s \cdot y \geq 1 \\ 1 - s \cdot y & \text{if } s \cdot y < 1 \end{cases}$$



Final form of CSVM

$$\min_{\tilde{w}, \tilde{b}, \{\xi_i\}} \frac{1}{2} \|w^*\|_2^2 + C \sum_{i=1}^{n} \xi_i$$

$$\text{st } y^i \cdot (\tilde{w}^T x^i + \tilde{b}) \geq 1 - \xi_i \quad \forall \; i \in 1 \ldots n$$

$$\& \quad \xi_i \geq 0 \quad \forall \; i = 1 \ldots n$$

$$\min_{\tilde{w}, \tilde{b}} \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{n} l_{hinge}(\tilde{w}^T x^i + \tilde{b}, y^i)$$

Corollary of Taylor's th:

$$f(x+\Delta x) \simeq f(x) + \Delta x \cdot f'(x) \quad \text{, if } \Delta x \text{ is "small"}$$

Taylor's th. in higher dim: (if we move along the vector $t = (t_1, \dots, t_d)$

then $f(x+t) \simeq f(x) + \sum_{i=1}^{d} t_i \cdot \frac{\partial f(x)}{\partial x_i}$

$$\boxed{f(x+t) = f(x) + t^T \nabla f(x)} \quad \text{if } t \text{ is "small"}$$

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_1 \partial x_3} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & & & & \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & - & - & - & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix}$$

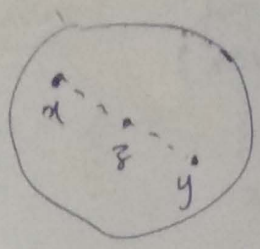
→ Hessian : $d \times d$ mat
: symmetric

If $\nabla f = 0$ & $\nabla^2 f$ is a PSD mat $\Rightarrow$ local/global min.

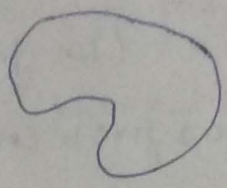
$\nabla f = 0$ & $\nabla^2 f$ is a NSD mat $\Rightarrow$ " " max

→ $x^T (\nabla^2 f) x \leq 0$

Convex Sets

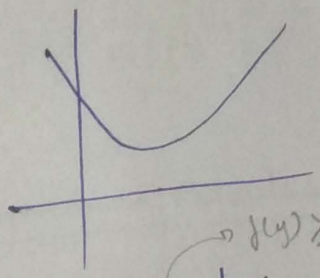


Convex set

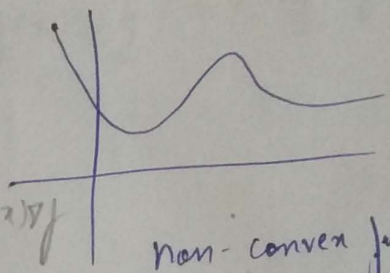


non-convex set

Convex set $\begin{cases} C \subseteq \mathbb{R}^d, \ \forall x, y \in C \ \& \ \forall \lambda \in [0,1] \\ z = \lambda \cdot x + (1-\lambda) \cdot y \in C \end{cases}$



convex fn.



non-convex fn.

$\forall x, y, \ \forall \lambda \in [0,1]$

$z = \lambda \cdot x + (1-\lambda) \cdot y$

$f(z) \leq \lambda \cdot f(x) + (1-\lambda) \cdot f(y)$

Convex fn.
- → must lie below all its chords.
- → for diff. fn, it must lie above all its tangents
- → for twice diff. fn, ∇^2 must be PSD everywhere

→ tang. to f at x_0 is the hyperplane.

$\nabla f(x_0)^T (x - x_0) + f(x_0) = 0$

→ sum of 2 convex fn.s is always convex

Eg $f(x) = 0$

$f(x) = c$

$f(x) = a^T x$ (lin. fn are convex)

$c \cdot f(x)$, if $c > 0$ & $f(x)$ is convex

if $g: \mathbb{R}^d \to \mathbb{R}$ is convex

$f: \mathbb{R} \to \mathbb{R}$ is convex & non dec.

then $f \circ g : \mathbb{R}^d \to \mathbb{R}$ is convex

$f(x) = \|x\|_2$ is convex

Subgradient - for convex, non-diff (possibly) f^n^s

for a given $f: \mathbb{R}^d \to \mathbb{R}$, possibly non-diff but convex

& x°

any vector g is ka the subgrad of f at x° if

$$f(x) \geq g^T(x - x^\circ) + f(x^\circ) \quad \forall x$$

Subdifferential - set of all subgrad of f at x°, denoted by $\partial f(x^\circ)$

$$\partial f(x^\circ) \triangleq \{ g : f(x) \geq g^T(x - x^\circ) + f(x^\circ), \forall x \}$$

2/12/19

Subgradient Calculus

→ $\partial(c \cdot f)(x) = c \cdot \partial f(x)$

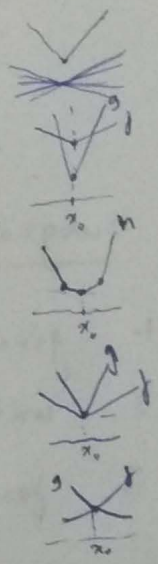
$= \{ c \cdot v : v \in \partial f(x) \}$

→ $\partial(f+g)(x) = \partial f + \partial g$

$= \{ u + v : u \in \partial f, v \in \partial g \}$

→ $\partial f(a^T x + b) = \partial f(a^T x + b) \cdot a$

$= \{ c \cdot a : c \in \partial f(a^T x + b) \}$



→ $h(x) = \max \{ f(x), g(x) \}$

?? if $f(x^\circ) > g(x^\circ)$ $\partial h(x^\circ) = \partial f(x^\circ)$

?? if $f(x^\circ) = g(x^\circ)$ $\partial h(x^\circ) = \{ \lambda u + (1-\lambda) v : u \in \partial f(x^\circ), v \in \partial g(x^\circ), \lambda \in [0,1] \}$

→ Stationary pt at x° if $0 \in \partial f(x^\circ)$. Local max/min is stationary even for non-diff. f^n^s

Subgradient for hinge loss

$$\ell_{hinge}(x) = \max\{1-x, 0\} = \max\{f(x), g(x)\}$$

$$\partial \ell_{hinge}(x) = \ell'_{hinge}(x) \quad \text{if } x \neq 1$$

At $x = 1$

$f(x) = 1 - x$ (diff-able) $\Rightarrow \partial f = f'(x) = -1$

$g(x) = 0$ (") $\Rightarrow \partial g = g'(x) = 0$

$$\Rightarrow \partial \ell_{hinge}(x) = \{-\lambda + (1-\lambda)0 : \lambda \in [0,1]\}$$
$$= [-1, 0]$$

$$\ell_{hinge}(y^i, \langle w, x^i \rangle) = [1 - y^i \langle w, x^i \rangle]_+$$

Finding $v^i \in \partial \ell_{hinge}(y^i, \langle w, x^i \rangle)$

$$v^i = \begin{cases} 0 & \text{if } y^i \langle w, x^i \rangle > 1 \\ -y^i \cdot x^i & \text{if } y^i \langle w, x^i \rangle < 1 \\ c \cdot y^i \cdot x^i & \text{if } y^i \langle w, x^i \rangle = 1 \\ c \in [-1, 0] & \end{cases}$$

Subgradient descent

1. given :- $f : \mathbb{R}^d \to \mathbb{R}$ to minimize
2. initialize $w^0 \in \mathbb{R}^d$
3. for $t = 0, 1 \ldots$
   1. obtain subgrad. $g^t \in \partial f(w^t)$
   2. Choose step length η_t → aka learning rate
   3. update $w^{t+1} \leftarrow w^t - \eta_t \cdot g^t$
   4. repeat until convergence.

For "nicely behaved" convex fn·s

set $\eta_t = \eta/\sqrt{t}$ → η: hyperparameter

or $\eta_t = \eta/t$

Armijo Rule : try some η_t, if not "nice", ↓ η_t

Adagrad : use a diff η for each dim. of w

η_t repl. by a diag. mat E^t i.e $w^{t+1} \leftarrow w^t - E^t g^t$

Adam : ~~used~~ uses momentum methods (injuses prev. grad. into the curr. grad)

Convergence → $\|w^{t+1} - w^t\| \to 0$

↳ general heuristics

if $\|g^t\|_2$ has become too small

if $|f(w^{t+1}) - f(w^t)|$ " " " "

if $f(w^t)$ is small enough

L8

$$J(w) = \frac{1}{2}\|w\|_2^2 + C \cdot \sum_{i=1}^{n} [1 - y^i \cdot w^T x^i]_+ \quad \text{(ignoring bias)}$$

?? $$\nabla J(w) = w + C \cdot \sum_{i=1}^{n} g^i y^i \cdot x^i \;, \quad g^i \in \nabla \ell_{hinge}(y^i \cdot w^T x^i)$$

$$w^{new} = w - \eta \cdot \nabla J(w)$$

$$= (1-\eta) w - \eta C \cdot \sum_{i=1}^{n} g^i y^i \cdot x^i$$

CSVM: $\min_{w,b} \frac{1}{2}\|w\|_2^2 + C \sum_{i=1}^{n} \ell_{hinge}(w^T x^i + b^i, y^i)$

$y^i \cdot (w^T x^i + b^i) \geq 1 - \xi_i \rightarrow \xi_i \geq 0$

Say if $n = 1$

$$w^{new} = (1-\eta) \cdot w - \eta C \cdot g^1 y^1 \cdot x^1$$

if small η : not much change in w

if large η : w changes as gradient dictates

if w does well on (x^1, y^1)

say $y^1 \cdot w^T x^1 > 1$, then $g^1 = 0$

$\Rightarrow$ no change in w due to pt (x^1, y^1)

if w does badly on (x^1, y^1)

say $y^1 \cdot w^T x^1 < 0$, then $g^1 = -1$

$\Rightarrow w^{new} = (1-\eta) \cdot w + \eta C y^1 \cdot x^1$

$\Rightarrow y^1 \cdot (w^{new})^T \cdot x^1$

$$= (1-\eta) y^1 \cdot w^T x^1 + \eta C \cdot \|x^1\|_2^2$$

→ ~~Model~~ GD actively tries to make the model perform better on all data pt.s.

Stochastic gradient method

$$\nabla J(w) = w + C \sum_{i=1}^{n} g^i y^i \cdot x^i \quad , \quad g^i \in \nabla l_{hinge}(y^i \cdot w^T x^i)$$

Calculating each g^i takes $O(d)$ time, thus total $O(nd)$ time

At each time, choose a random data pt. (x^{i_t}, y^{i_t})

$$\nabla J(w) \simeq w + C \cdot g^{i_t} y^{i_t} \cdot x^{i_t} \quad - \text{ only } O(d) \text{ time}$$

Mini-batch SGD. – if data is diverse, stoc. grad. may vary quite a lot

Choose B random data pt.s, say $(x^{i_t^1}, y^{i_t^1}) \cdots (x^{i_t^B}, y^{i_t^B})$

↳ mini batch size

$$\nabla J(w) \simeq w + C \cdot \sum_{b=1}^{B} g^{i_t^b} y^{i_t^b} x^{i_t^b}$$

↳ $O(Bd)$ time

if $B = n$, MBSGD becomes GD.

Constrained optimization

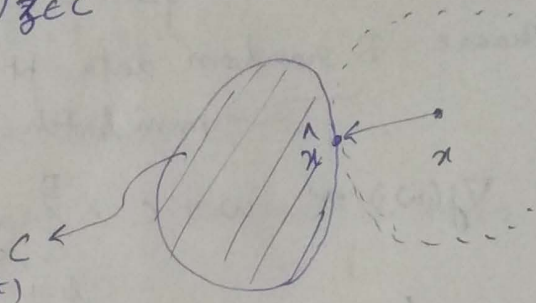
?? Method 1 : Interior pt method.

Method 2 : Projected grad. desc

→ perf. SGD as usual. However, if this causes us to go out of feasible set C, go back to feasible set.

going back : projection step

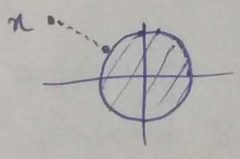
$$\hat{x} = \Pi_C(x) = \arg\min_{z \in C} \|x - z\|_2^2$$



1. Choose $g^t \in \partial f(w^t)$
2. $u^{t+1} \leftarrow w^t - \eta_t \cdot g^t$
3. $w^{t+1} \leftarrow \Pi_C(u^{t+1})$
4. Repeat until convergence

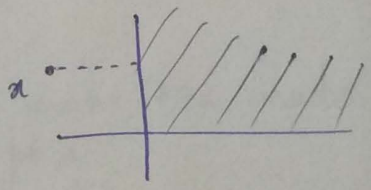
Useful projections $\hat{x} = \Pi_C(x)$

$C = \{x : \|x\|_2 \leq 1\}$



$$\hat{x} = \begin{cases} x & \text{if } \|x\|_2 \leq 1 \\ \frac{x}{\|x\|_2} & \text{if } \|x\|_2 > 1 \end{cases}$$

$C = \{x : x_i \geq 0\}$

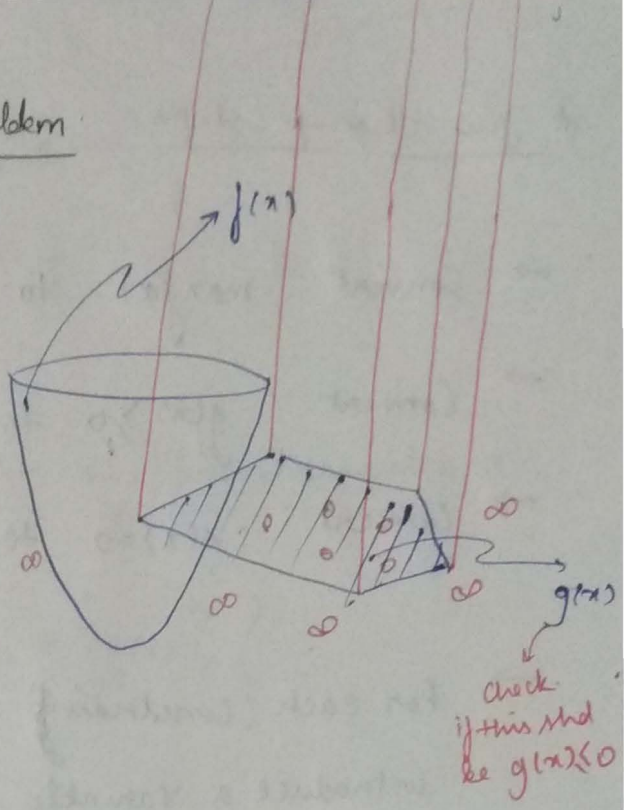


$$\hat{x}_i = \begin{cases} x_i & \text{if } x_i \geq 0 \\ 0 & \text{if } x_i < 0 \end{cases}$$

Method 3 : Creating a dual problem

Suppose we wish to solve

$$\min_{x \in \mathbb{R}^d} f(x)$$
$$\text{s.t } g(x) \leq 0$$



Construct a barrier fn. $r(x)$

s.t $r(x) = 0$ if $g(x) \leq 0$

$r(x) = \infty$ otherwise

and simply solve

$$\min_{x \in \mathbb{R}^d} f(x) + r(x)$$

One way to construct such a barrier

$$r(x) = \max_{\alpha \geq 0} \alpha \cdot g(x)$$

Thus we want to solve

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) + \max_{\alpha \geq 0} \alpha \cdot g(x) \right\}$$

same as

$$\min_{x \in \mathbb{R}^d} \left\{ \max_{\alpha \geq 0} \left\{ f(x) + \alpha \cdot g(x) \right\} \right\}$$

4/12/19

A few cleanup steps

→ Convert $\max f(x)$ to $\min \{-f(x)\}$

→ Convert $g(x) \geq 0$ to $-g(x) \leq 0$

→ Convert $s(x) = 0$ to $s(x) \leq 0$
& $-s(x) \leq 0$

→ for each constraint $g_i(x) \leq 0$ $i = 1 \cdots c$
introduce a variable α_i $i = 1 \cdots c$
↳ dual var.s
or Lagrangian multipliers

5/12/19

The Lagrangian

$$\min_x f(x)$$
$$\text{st } g_1(x) \leq 0$$
$$g_2(x) \leq 0$$
$$\vdots$$
$$g_c(x) \leq 0$$

$$\mathcal{L}(x, \alpha) = f(x) + \sum_{c=1}^{C} \alpha_c \cdot g_c(x)$$

↳ Lagrangian.

If x violates even one constraint, we have

$$\max_{\alpha \in \mathbb{R}^C} \{\mathcal{L}(x, \alpha)\} = \infty$$
$$\alpha_c \geq 0$$

If x satisfies every single constraint, we have

$$\max_{\alpha \in \mathbb{R}^C} \{\mathcal{L}(x, \alpha)\} = f(x)$$
$$\alpha_c \geq 0$$

$$\min_{x \in \mathbb{R}^d} \left\{ \max_{\substack{\alpha \in \mathbb{R}^C \\ \alpha_c \geq 0}} \left\{ f(x) + \sum_{c=1}^{C} \alpha_c \cdot g_c(x) \right\} \right\}$$

The dual prob.

original prob :- primal problem w/ primal variables x.

$$\min_{x \in \mathbb{R}^d} \left\{ \max_{\substack{\alpha \in \mathbb{R}^C \\ \alpha_c \geq 0}} \left\{ f(x) + \sum_{c=1}^{C} \alpha_c \cdot g_c(x) \right\} \right\}$$

dual prob

??

$$\max_{\substack{\alpha \in \mathbb{R}^C \\ \alpha_c \geq 0}} \left\{ \min_{x \in \mathbb{R}^d} \left\{ f(x) + \sum_{c=1}^{C} \alpha_c \cdot g_c(x) \right\} \right\}$$

let $\hat{x}^P, \hat{\alpha}^P$ be the soln to primal prob.

$\hat{x}^D, \hat{\alpha}^D$ " " dual prob.

Strong duality : $\hat{x}^P = \hat{x}^D$ if the orig prob. is convex & "nice"

Complementary slackness : $\hat{\alpha}_c^D \cdot g_c(\hat{x}^D) = 0 \quad \forall$ constraints c

SVM max $\left\{ \min_{i=1:n} \frac{|w^T x^i + b|}{\|w\|_2} \right\}$ s.t. $y^i (w^T x^i + b) > 0$

Hard SVM w/out bias

$$\min_{w \in \mathbb{R}^d} \frac{1}{2} \|w\|_2^2 \quad \text{s.t.} \quad 1 - y^i \cdot w^T x^i \leq 0 \quad \forall \; i = 1 \dots n$$

↳ n constraints ⇒ n dual var.
⇒ $\alpha \in \mathbb{R}^n$

$$\mathcal{L}(w, \alpha) = \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y^i w^T x^i)$$

?? Primal: $\underset{w \in \mathbb{R}^d}{\text{argmin}} \left\{ \underset{\alpha \geq 0}{\text{argmax}} \left\{ \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y^i w^T x^i) \right\} \right\}$

?? Dual: $\underset{\alpha \geq 0}{\text{argmax}} \left\{ \underset{w \in \mathbb{R}^d}{\text{argmin}} \left\{ \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y^i w^T x^i) \right\} \right\}$

Simplifying the dual problem

The inner is the dual problem

∵ This an unconstrained prob. w/ a convex & diff. able objective, use first order optimality

We get $\boxed{w = \sum_{i=1}^{n} \alpha_i y^i \cdot x^i}$

Subs. back in dual prob.

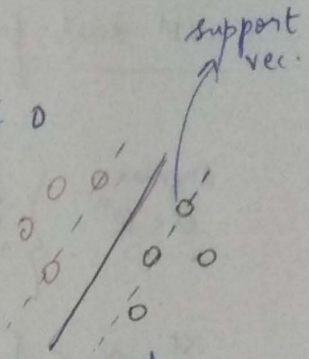
$$\underset{\alpha \geq 0}{\text{argmax}} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^i y^j \langle x^i, x^j \rangle \right\}$$

Once the optimal val of α is known, use

$$w = \sum_{i=1}^{n} \alpha_i y^i \cdot x^i$$

Support Vectors

α_i ∀ data pt.s



Support vec :- data pt.'s for which $\alpha_i \neq 0$

Using complementary slackness

$$\alpha_i \left(1 - y^i \omega^T x^i\right) = 0$$

$\Rightarrow$ only those data pt's can become SV for which

$y^i \omega^T x^i = 1 \Rightarrow$ at margin

Dual for CSVM

if we have bias b as well as slack ξ then dual is given by

$$\underset{\alpha \in \mathbb{R}^n}{\text{argmax}} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^i y^j \langle x^i, x^j \rangle \right\}$$

$$\text{st. } \alpha_i \in [0, C_i] \;\&\; \sum_{i=1}^{n} \alpha_i y^i = 0$$

Analogy

α^i : force on hyperplane ω by i^{th} data pt in y^i dir.

$\xi_i = 0$ as $\sum_{i=1}^{n} \alpha_i y^i = 0$

Also, since $\omega = \sum_{i=1}^{n} \alpha_i y^i x^i$

$\xi_i = 0$

Thus the support vec. mechanically support the hyperplane.

6/12/19

CSVM dual problem

$$\underset{\alpha \in \mathbb{R}^n}{\text{argmax}} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^i y^j \langle x^i, x^j \rangle \right\}$$

st $\alpha_i \in [0, C_i]$ & $\underbrace{\sum_{i=1}^{n} \alpha_i y^i = 0}$

if bias is present

Coordinate Descent (popularly applied to dual probs.)

(SDCA - stochastic dual co-ord ascent)

Only one co-ord changed in a single step.

∵ dual prob is a maximiz. pr

Eg $\min_{x \in \mathbb{R}^d} f(x)$ st $x \in C$

$$\nabla_j f(x) = \frac{\partial f}{\partial x_j}$$

CCD : cyclic order

$j_t = 1, 2 \ldots d, 1, \ldots, d \ldots$

SCD : randomly chosen

Block CD : choose a set of co-ord's to be updated at each t

Projected Co-ord descent

1. For $t = 0, 1 \cdots$
   1. Select a coord $j_t \in [d]$
   2. Let $u_{j_t}^{t+1} \leftarrow x_{j_t}^t - \eta \nabla_{j_t} f(x^t)$
   3. Let $u_j^{t+1} \leftarrow x_j^t$ for $j \neq j_t$
   4. Project $x^{t+1} \leftarrow \pi_C(u^{t+1})$
   5. Repeat until convergence