① **Spring 2023**

Linear Classifiers and Perceptrons.

Note Title — 11/12/2006

N samples: $\{(\underline{x}_\mu, \omega_\mu) : \mu = 1 \text{ to } N\}$,
$\omega_\mu \in \{\pm 1\}$,

Can we find a linear classifier that separates the positive and negative examples?

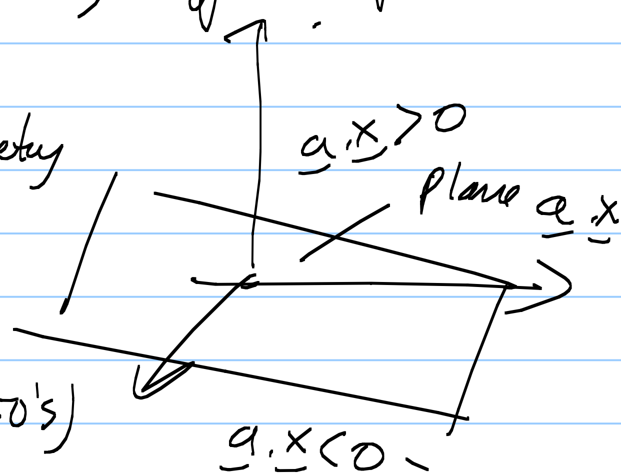
EG. a plane $\underline{a}\cdot\underline{x} = 0$ s.t. $\text{sign}(\underline{a}\cdot\underline{x}) = \omega$

s.t. $\underline{a}\cdot\underline{x}_\mu \geq 0$, if $\omega_\mu = +1$

$\underline{a}\cdot\underline{x}_\mu \leq 0$, if $\omega_\mu = -1$.

Plane goes through the origin $(\underline{a}\cdot\underline{0} = 0)$

Geometry



Perceptron Algorithm (1950's)

First, replace -ve examples by +ve examples

if $\omega_\mu = -1$, set $\underline{x}_\mu \to -\underline{x}_\mu$, $\omega_\mu \to -\omega_\mu$.

(Note. require $\text{sign}(\underline{a}\cdot\underline{x}_\mu) = \omega_\mu$, this is equivalent to $\text{sign}(-\underline{a}\cdot\underline{x}_\mu) = -\omega_\mu$)

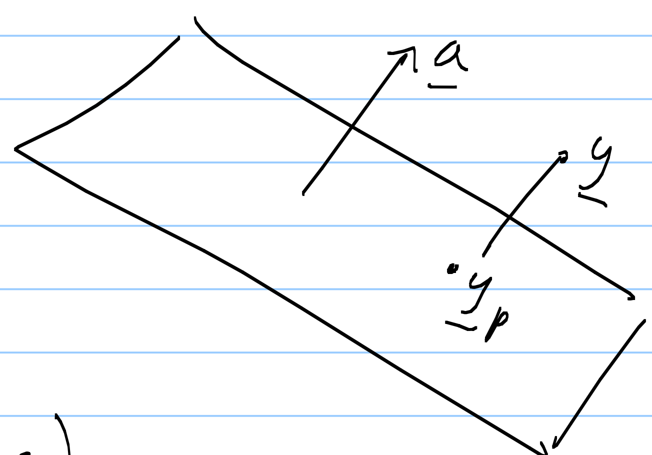
Spring 2017.

(2) This reduces to finding a plane s.t. $\underline{a} \cdot \underline{x}_\mu \geq 0$, for $\mu = 1$ to N

__Note:__ the vector $\underline{a}$ need not be unique. It is better to try to maximize the margin (see next lecture). To find $\underline{a}$ with $|\underline{a}| = 1$, so that $\underline{a} \cdot \underline{x}_\mu \geq m$, $\forall \mu = 1$ to N for the maximum value of m.

More geometry

__Claim:__



If $\underline{a}$ is a unit vector $|\underline{a}| = 1$, then $\underline{a} \cdot \underline{y}$ is the sign(*) distance of $\underline{y}$ to the plane $\underline{a} \cdot \underline{x} = 0$.

(* i.e. $\underline{a} \cdot \underline{y} > 0$, if $\underline{y}$ is above plane
$\underline{a} \cdot \underline{y} < 0$, if $\underline{y}$ is below plane)

__Proof__ write $\underline{y} = \lambda \underline{a} + \underline{y}_p$, where $\underline{y}_p$ is the projection of $\underline{y}$ into the plane. By definition $\underline{a} \cdot \underline{y}_p = 0$, hence $\lambda = (\underline{a} \cdot \underline{y}) / (\underline{a} \cdot \underline{a}) = (\underline{a} \cdot \underline{y})$, if $|\underline{a}| = 1$.

Spring 2013

(3) **Perceptron Algorithm.**

Initialize: $\underline{a}(0) = 0$.

Loop over $\mu = 1$ to N

if $\underline{x}_\mu$ is missclassified, set $\underline{a} \to \underline{a} + \underline{x}_\mu$,

Repeat until all samples are classified correctly.

Novikov's Thm. The Perceptron algorithm will converge to a solution weight that classifies all the samples correctly (provided this is possible).

Proof. Let $\hat{\underline{a}}$ be a separating weight

Let $m = \min_{\mu=1}^{N} \hat{\underline{a}} \cdot \underline{x}_\mu \qquad (m>0)$

Let $\beta^2 = \max_{\mu=1}^{N} |\underline{x}_\mu|^2$

Suppose $\underline{x}_t$ is missclassified at ~~time~~ t

so $\underline{a} \cdot \underline{x}_t < 0$

$$\underline{a}_{t+1} - (\beta^2/m)\,\hat{\underline{a}} = \underline{a}_t - (\beta^2/m)\,\hat{\underline{a}} + \underline{x}_t$$

Spring 2013

(4) $$\|\underline{a}_{t+1} - \beta^2/m\,\hat{\underline{a}}\|^2 = \|\underline{a}_t - (\beta^2/m)\,\hat{\underline{a}}\|^2 + \|\underline{x}_t\|^2 - 2(\underline{a}_t - (\beta^2/m)\hat{\underline{a}})\cdot\underline{x}_t.$$

Using $\|\underline{x}_t\|^2 \leq \beta^2$, $\underline{a}_t\cdot\underline{x}_t < 0$, $-\hat{\underline{a}}\cdot\underline{x}_t < -m$

it follows that

$$\|\underline{a}_{t+1} - \beta^2/m\,\hat{\underline{a}}\|^2 \leq \|\underline{a}_t - \beta^2/m\,\hat{\underline{a}}\|^2 + \beta^2 - 2\beta^2/m \cdot m$$

Hence $$|\underline{a}_{t+1} - \beta^2/m\,\hat{\underline{a}}\|^2 \leq |\underline{a}_t - \beta^2/m\,\hat{\underline{a}}\|^2 - \beta^2.$$

So, each time we update a weight, we reduce the quantity $\|\underline{a}_t - \beta^2/m\,\hat{\underline{a}}\|^2$ by a fixed amount β^2. $\|\underline{a}_0 - \beta^2/m\,\hat{\underline{a}}\|^2$ is bounded by $\frac{\beta^4}{m^2}|\hat{\underline{a}}|^2$.

So we can update the weights at most $\frac{\beta^2}{m^2}|\hat{\underline{a}}|^2$ times

Guarantees Convergence.

(5) SVM's Spring 2013

Linear Separation: Margins & Duality

Note Title 11/12/2006

Modern approach to linear separation.

Data $\{(\underline{x}_\mu, \omega_\mu) : \mu = 1 \text{ to } N\}$, $\omega_\mu \in \{-1, 1\}$

Hyperplane $\{\underline{x} : \underline{x}\cdot\underline{a} + b = 0\}$, $|\underline{a}| = 1$.

The signed distance of a point $\underline{x}$ to the plane is $\underline{a}\cdot\underline{x} + b$.

Line $\underline{x}(\lambda) = \underline{x} + \lambda\underline{a}$ ← to project on plane

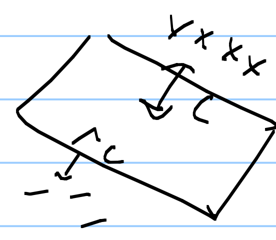
Hits plane when $\underline{a}\cdot(\underline{x} + \lambda\underline{a}) = -b$

$\lambda = -(\underline{a}\cdot\underline{x} + b)/|\underline{a}|^2 = -(\underline{a}\cdot\underline{x} + b)$ if $|\underline{a}| = 1$.

Seek classifier with biggest margin

$$\max_{a, b, |a|=1} C \quad \text{s.t.} \quad y_\mu(\underline{x}_\mu\cdot\underline{a} + b) \geq C, \quad \forall \mu = 1 \text{ to } N.$$

i.e. the positive examples are at least distance C above the plane, and negative examples are at least C below the plane.



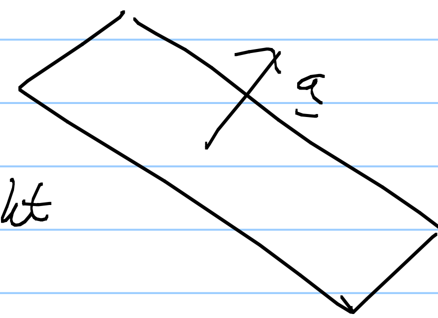
Large margin is good for generalization (less chance of an accidental alignment)

(6)

Spring 2023

Now, allow for some datapoints to be missclassified.

Slack variables → allow datapoints to move in direction $\underline{a}$, so that they are on the right side of the margin.



slack variables $\{z_1, \ldots z_N\}$

Criterion: $\max\limits_{a, b, |\underline{a}|=1} C$ s.t. $y_\mu(\underline{x}_\mu \cdot \underline{a} + b) \geq C(1 - z_\mu)$ $\forall \mu \in \{1, N\}$

with constraint $z_\mu \geq 0$, $\forall \mu$.

Alternatively $y_\mu\{(\underline{x}_\mu + C z_\mu \underline{a}) \cdot \underline{a} + b\} \geq C$

like moving $\underline{x}_\mu$ to $\underline{x}_\mu + C z_\mu \underline{a}$.

But, you must pay a penalty for using slack variables. A penalty like $\sum\limits_{\mu=1}^{N} z_\mu$.

If $z_\mu = 0$, then the datapoint is correctly classified and is past the margin.

If $z_\mu > 0$, then the datapoint is on the wrong side of the margin

(7) Spring 2017

Task: We need to estimate several quantities simultaneously:

(1) The plane $\underline{a}, b$

(2) The margin C

(3) The slack variables $\{z_\mu\}$.

We need a criterion that maximizes the margin and minimizes the amount of slack variables used.

Remove criterion $|\underline{a}|=1$, set $C = 1/|\underline{a}|$.

Criterion: Min $\frac{1}{2}\underline{a}.\underline{a} + \gamma \sum_\mu z_\mu$

s.t. $y_\mu(\underline{x}_\mu.\underline{a}+b) \geq 1 - z_\mu, \quad \forall \mu$

$z_\mu \geq 0, \forall \mu.$

Quadratic Primal Problem requires Lagrange multipliers.

$$L_p = \frac{1}{2}\underline{a}.\underline{a} + \gamma \sum_\mu z_\mu - \sum_\mu \alpha_\mu \{y_\mu(\underline{x}_\mu.\underline{a}+b) - (1-z_\mu)\} - \sum_\mu \tau_\mu z_\mu.$$

The $\{\alpha_\mu\}$ & $\{\tau_\mu\}$ are Lagrange parameters needed to enforce the inequality constraints. $\alpha_\mu \geq 0, \tau_\mu \geq 0, \forall \mu.$

(8)

Spring 2013.

L_p is a function of the primal variables $\underline{a}, b, \{z_\mu\}$ and the lagrange parameters $\{\alpha_\mu, \tau_\mu\}$

There is no analytic solution for these variables, but we can use analytic techniques to get some understanding of their properties.

$$\frac{\partial L_p}{\partial \underline{a}} = 0 \quad \Rightarrow \quad \hat{\underline{a}} = \sum_\mu \hat{\alpha}_\mu y_\mu \underline{x}_\mu .$$

$$\frac{\partial L_p}{\partial b} = 0 \quad \Rightarrow \quad \sum_\mu \hat{\alpha}_\mu y_\mu = 0.$$

$$\frac{\partial L_p}{\partial z_\mu} = 0 \quad \Rightarrow \quad \hat{\alpha}_\mu = \gamma - \hat{\tau}_\mu \,, \quad \forall \mu.$$

The classifier is

$$\mathrm{sign}\{ \hat{\underline{a}} \cdot \underline{x} + \hat{b} \} = \mathrm{sign}\{ \sum_\mu \alpha_\mu y_\mu \underline{x}_\mu \cdot \underline{x} + \hat{b} \}$$

Support vectors, the solution depends only on the vectors $\underline{x}_\mu$ for which $\alpha_\mu \neq 0$. //

(9)

Spring 2013.

The constraints are

$$y_\mu(\underline{x}_\mu \cdot \hat{\underline{a}} + \hat{b}) \geq 1 - \hat{z}_\mu$$

$$\hat{z}_\mu \geq 0, \quad \hat{\tau}_\mu \geq 0.$$

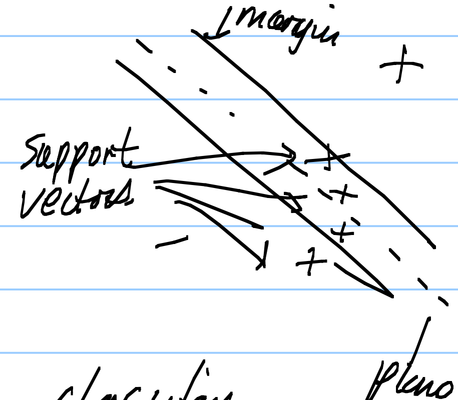
By theory of Quadratic Programming —

$\alpha_\mu > 0$, only if either:

(i) $z_\mu > 0$ slack variable is used

(ii) $z_\mu = 0$, but $y_\mu(\underline{x}_\mu \cdot \hat{\underline{a}} + \hat{b}) = 1$
datapoint is on the margin

The classifier depends only on the support vectors, the other datapoints do not matter.



This is intuitively reasonable — the classifier must pay close attention to the data that is difficult to classify — the data near the boundary

This differs from the probabilistic approach — where we learn probability models for each class and then use the Bayes classifier.

(10)

Spring 2013.

Dual Formulation

We can solve the problem more easily in the dual formulation. — function of Lagrange multipliers only.

$$L_d = \sum_\mu \alpha_\mu - \frac{1}{2} \sum_{\mu,\nu} \alpha_\mu \alpha_\nu y_\mu y_\nu \underline{x}_\mu \cdot \underline{x}_\nu .$$

with constraint $0 \le \alpha_\mu \le \tau$, $\sum_\mu \alpha_\mu y_\mu = 0$.

There are standard packages to solve this.

Knowing $\{\hat{\alpha}_\mu\}$, will give us the solution

$\hat{\underline{a}} = \sum_\mu \hat{\alpha}_\mu y_\mu \underline{x}_\mu$, (only a little more work to get $\hat{b}$)

(11)

Spring 2013

Relationship between Primal & Dual.

Start with dual formulation. L_p

Rewrite it as

$$L_p = \left(-\tfrac{1}{2}\right) \underline{a}\cdot\underline{a} + \sum_\mu \alpha_\mu + \underline{a}\cdot\left\{\underline{a} - \sum_\mu \alpha_\mu y_\mu \underline{x}_\mu\right\}$$

$$+ \sum_\mu z_\mu \left\{\gamma - \tau_\mu - \alpha_\mu\right\} - b \sum_\mu \alpha_\mu y_\mu .$$

Extremize w.r.t. $\underline{a}$, b, $\{z^\mu\}$ gives:

$$\hat{\underline{a}} = \sum_\mu \alpha_\mu y_\mu \underline{x}_\mu, \quad \sum_\mu \alpha_\mu y_\mu = 0, \quad \gamma - \tau_\mu - \alpha_\mu = 0$$

Substituting back into L_p gives:

$$L_d = -\frac{1}{2} \sum_{\mu,\nu} \alpha_\mu \alpha_\nu y_\mu y_\nu \underline{x}_\mu \cdot \underline{x}_\nu + \sum_\mu \alpha_\mu$$

maximize w.r.t. $\{\alpha_\mu\}$.

(12) Spring 2013

The Perceptron can be reformulated in this way.

By the theory, the weight hypothesis will always be of form:

$$\underline{a} = \sum_\mu \alpha_\mu y_\mu \underline{x}_\mu.$$

Perceptron Update Rule:

If data $\underline{x}_\mu$ is missclassified

i.e. $y_\mu(\underline{a}.\underline{x}_\mu + b) \leq 0$

Set $\alpha_\mu \rightarrow \alpha_\mu + 1$

$b \rightarrow b + y_\mu R^2.$

R is radius of smallest ball containing the data.

(13) Spring 2013

$$L_p = \frac{1}{2}\underline{a}\cdot\underline{a} + \gamma \sum_\mu z_\mu$$

constraints.

$y_\mu(\underline{x}_\mu\cdot\underline{a}+b) - (1-z_\mu) \geq 0$

$z_\mu \geq 0$

$z_\mu \geq 1 - y_\mu(\underline{x}_\mu\cdot\underline{a}+b)$

Hence $L_p = \frac{1}{2}\underline{a}\cdot\underline{a} + \gamma \sum_\mu \max\{0, 1-y_\mu(\underline{x}_\mu\cdot\underline{a}+b)\}$

Hinge Loss – Loss = 0 if $y_\mu(\underline{x}_\mu\cdot\underline{a}+b) \geq 1$

i.e. data point is on the correct side of margin

loss $= 1-y_\mu(\underline{x}_\mu\cdot\underline{a}+b)$

amount of slack variable required to move datapoint to correct side of margin.

(Recall – empirical loss + regularizer)

Online Learning:

At time t,

select datapoint $(\underline{x}_\mu, y_\mu)$

$\underline{a}^t \to \underline{a}^t - \Delta \frac{\partial L_p}{\partial \underline{a}}$ one iteration of steepest descent

repeat.

$\frac{\partial L_p}{\partial \underline{a}} = \underline{a} + 0$, if $y_\mu(\underline{x}_\mu\cdot\underline{a}+b) \geq 1$

data point is correctly classified

$\frac{\partial L_p}{\partial \underline{a}} = \underline{a} - \gamma\, y_\mu \underline{x}_\mu$

Update. $\underline{a}^{t+1} = \underline{a}^t - \Delta \underline{a}^t$, if $y_\mu(\underline{x}_\mu\cdot\underline{a}+b) \geq 1$

$= \underline{a}^t - \Delta \underline{a}^t + \Delta\gamma\, y_\mu \underline{x}_\mu$, otherwise

similar to perceptron.