What's the Perceptron Optimizing?

Machine Learning – CSE446
Carlos Guestrin
University of Washington

May 1, 2013

The Perceptron Algorithm [Rosenblatt '58, '62]

- Classification setting: y in {-1,+1}
- Linear model
  - Prediction: $\hat{y} = \text{Sign}(w \cdot x)$
- Training: $w^{(0)} = 0$ or something smarter
  - Initialize weight vector:
  - At each time step:
    - Observe features: $x^{(t)} \leftarrow$ (page, user, ad)
    - Make prediction: $\hat{y} = \text{sign}(w^{(t)} \cdot x^{(t)})$
    - Observe true class:
      $y^{(t)} \leftarrow$ true label
    - Update model:
      - If prediction is not equal to truth, if make a mistake!

if $\hat{y} = y^{(t)}$
$w^{(t+1)} \leftarrow w^{(t)}$
else
$w^{(t+1)} \leftarrow w^{(t)} + \underbrace{y^{(t)} x^{(t)}}_{\text{gradient}}$

$w^{(t+1)} \leftarrow w^{(t)} + 1\!\!1(\text{mistake})(y^{(t)} x^{(t)})$

I made a mistake:
e.g. $y^{(t)} = +1$
$w^{(t)} \cdot x^{(t)} < 0$
but wanted > 0
what u max $u \cdot x^{(t)}$? $x^{(t)}$!!
by adding $x^{(t)}$ to w
I increase $w^{(t+1)} \cdot x^{(t)}$ the most

similarly when $y^{(t)} = -1$

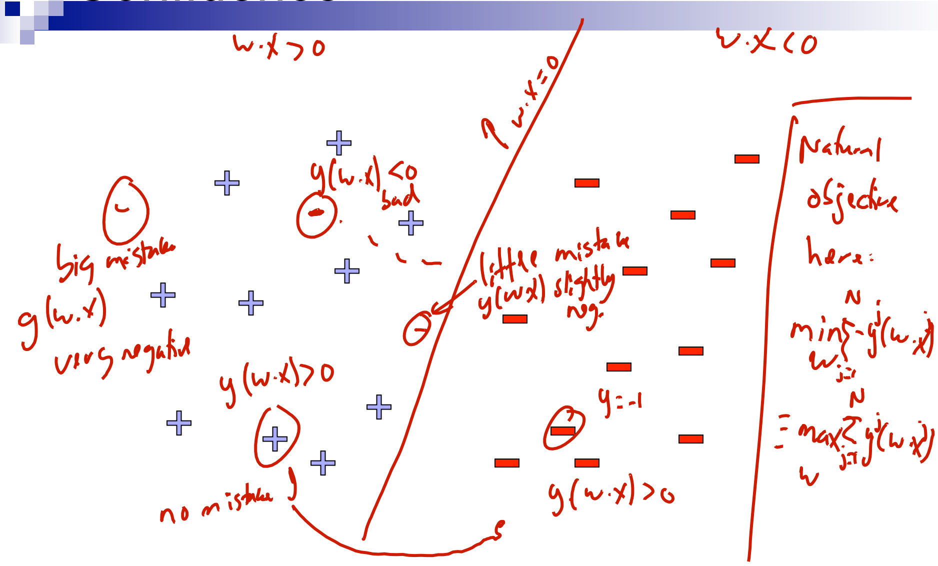
What is the Perceptron Doing???

- When we discussed logistic regression:
  - Started from maximizing conditional log-likelihood

$$\max_w P(Y|X,w)$$

- When we discussed the Perceptron:
  - Started from description of an algorithm
- What is the Perceptron optimizing????

Perceptron Prediction: Margin of Confidence

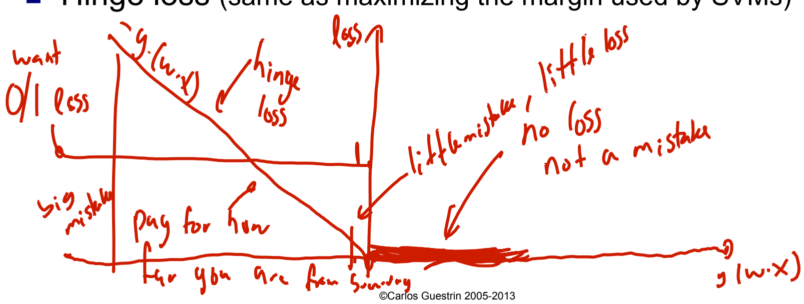


Hinge Loss

- Perceptron prediction: $\mathrm{Sign}(w \cdot x)$
- Makes a mistake when:

$y(w \cdot x) < 0 \implies \quad \ell(w,x) = \begin{cases} 0 & \text{if no mistake } y(w \cdot x) \geq 0 \\ -y(w \cdot x) & \text{otherwise } y(w \cdot x) < 0 \end{cases}$

- Hinge loss (same as maximizing the margin used by SVMs)



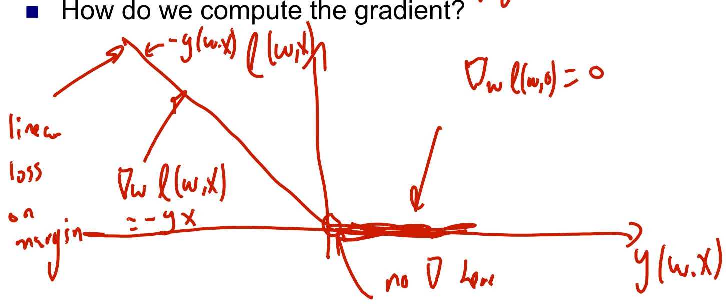
$(a)_+ \begin{cases} a & \text{if } a > 0 \\ 0 & \text{otherwise} \end{cases}$

Minimizing hinge loss in Batch Setting

- Given a dataset: $(x^1, y^1) \dots (x^N, y^N)$
- Minimize average hinge loss:

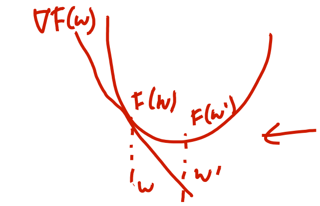
$$\min_w \frac{1}{N} \sum_{j=1}^{N} \ell(w,x) = \begin{cases} 0 & \text{if } y(w \cdot x) \geq 0 \\ -y(w \cdot x) & \text{otherwise} \end{cases} \Bigg\} \left(-y(w \cdot x)\right)_+$$

- How do we compute the gradient?



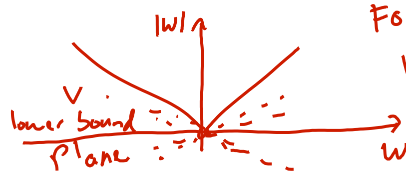
Subgradients of Convex Functions

- Gradients lower bound convex functions:



$F(w') \geq F(w) + \nabla F(w)^T (w'-w)$

- Gradients are unique at **w** iff function differentiable at **w**
- Subgradients: Generalize gradients to non-differentiable points:
  - Any plane that lower bounds function:



For $|w|$ at 0:

$v \in [-1,1]$

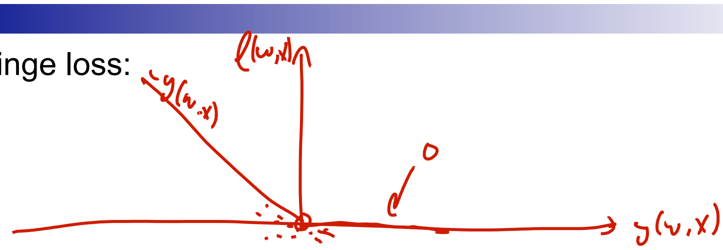
$v \in \partial_w F(w)$ subgradient

iff

$F(w') \geq F(w) + v^T(w'-w)$

Subgradient of Hinge

- Hinge loss:



- Subgradient of hinge loss:
  - If $y^{(t)} (w.\mathbf{x}^{(t)}) > 0$: $\partial \ell(w,x) = 0$
  - If $y^{(t)} (w.\mathbf{x}^{(t)}) < 0$: $\partial \ell(w,x) = -yx$
  - If $y^{(t)} (w.\mathbf{x}^{(t)}) = 0$: $\partial \ell(w,x) = [-yx, 0]$ e.g. $-yx$
  - In one line:

$$\partial \ell(w,x) = \underbrace{\mathbb{1}\left(y\,(w \cdot x) \leq 0\right)}_{\text{indicator of a mistake}} (-yx)$$

Subgradient Descent for Hinge Minimization

- Given data: $(x^1, y^1) \ldots (x^N, y^N)$

  want $\min_w$

- Want to minimize:

$$\frac{1}{N}\sum_{j=1}^{N} \ell(w, x^j) = \frac{1}{N}\sum_{j=1}^{N} \left(-y^j (w \cdot x^j)\right)_+$$

- Subgradient descent works the same as gradient descent:
  - But if there are multiple subgradients at a point, just pick (any) one:

$$w^{(t+1)} \leftarrow w^{(t)} - \underbrace{\eta}_{\text{step size}} \sum_{j=1}^{N} \underbrace{\partial \ell(w, x^j)}_{\mathbb{1}\left(\substack{\text{mistake:} \\ y^j (w \cdot x^j) \le 0}\right)(-y^j x^j)}$$

Perceptron Revisited

- Perceptron update:

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} + \overbrace{\mathbb{1}}^{\eta = 1}\underbrace{\left[y^{(t)}(\mathbf{w}^{(t)} \cdot \mathbf{x}^{(t)}) \le 0\right]}_{\text{if mistake}} \underbrace{y^{(t)}\mathbf{x}^{(t)}}_{\text{subgradient}}$$

- Batch hinge minimization update: sum over data point if mistake on point i

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} + \eta \frac{1}{N}\sum_{i=1}^{N}\left\{\mathbb{1}\left[y^{(i)}(\mathbf{w}^{(t)} \cdot \mathbf{x}^{(i)}) \le 0\right] \underbrace{y^{(i)}\mathbf{x}^{(i)}}_{\text{subgradient}}\right\}$$

param ... hopefully $\eta > 1$

- Difference? perceptron update is a stochastic gradient descent alg for hinge loss minimization with fixed step size $(\eta = 1)$

What you need to know

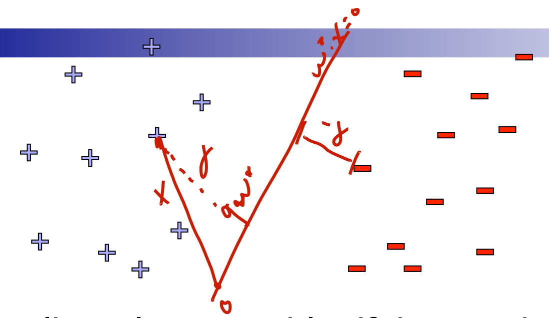
- Perceptron is optimizing hinge loss
- Subgradients and hinge loss
- (Sub)gradient decent for hinge objective

Kernels

Machine Learning – CSE446
Carlos Guestrin
University of Washington

May 1, 2013

Linear Separability: More formally, Using Margin



- Data linearly separable, if there exists
  - a vector $\exists w^*$, $\|w^*\|=1$
  - a margin $\gamma > 0$
- Such that all points are γ far or more from $w^* \cdot x = 0$

$\forall t$ if $y^{(t)} = +1$ $w^* \cdot x^{(t)} > \gamma$

$y^{(t)} = -1$ $w^* \cdot x^{(t)} < -\gamma$

$\underbrace{y^{(t)}\, w^* \cdot x^{(t)} > \gamma}_{\text{linearly separable, margin } \gamma}$

Perceptron Analysis: Linearly Separable Case

- Theorem [Block, Novikoff]:
  - Given a sequence of labeled examples: $(x^{(1)},y^{(1)}), (x^{(2)},y^{(2)}) \ldots (x^{(T)},y^{(T)})$

    examples need not be iid. or random..
  - Each feature vector has bounded norm:

    $\forall t \ \|x^{(t)}\| \le R$ w^* is unknown!
  - If dataset is linearly separable:

    $\exists w^*, \|w^*\|=1 \ \forall t \ y^{(t)}\, w^* \cdot x^{(t)} \geq \gamma$, for $\gamma > 0$
- Then the number of mistakes made by the online perceptron on ~~this~~ any sequence is bounded by

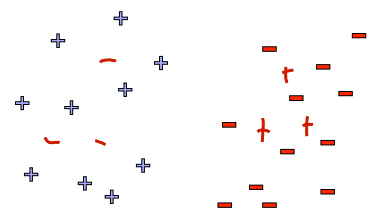
$$\left(\frac{R}{\gamma}\right)^2$$

wow!!

constant, doesn't depend on T or dimensionality of X !!

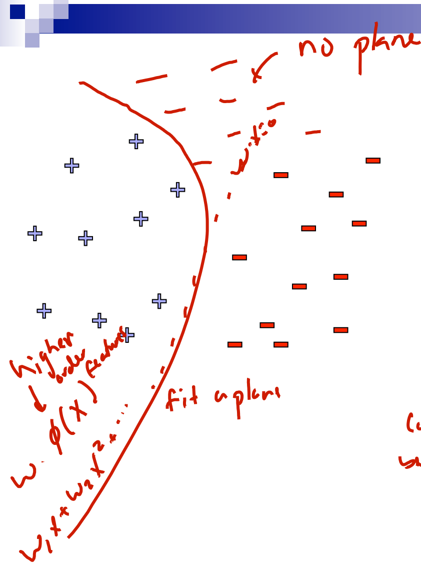
Beyond Linearly Separable Case

- Perceptron algorithm is super cool!
  - No assumption about data distribution!
    - Could be generated by an oblivious adversary, no need to be iid
  - Makes a fixed number of mistakes, and it's done for ever! $\leftarrow \left(\frac{R}{\gamma}\right)^2$
    - Even if you see infinite data
- However, real world not linearly separable
  - Can't expect never to make mistakes again
  - Analysis extends to non-linearly separable case
  - Very similar bound, see Freund & Schapire
  - Converges, but ultimately may not give good accuracy (make many many many mistakes)



we need features that make data as linearly separable as possible

What if the data is not linearly separable?



Use features of features of features of features….

$$\Phi(\mathbf{x}) : R^{m} \mapsto F$$

(handwritten: "k dim" over $\mathbf{x}$, "k" over m; "space here" beside F)

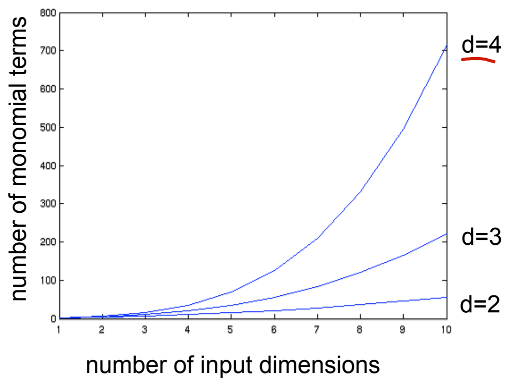
$$\phi(x) = \begin{bmatrix} x \\ x^2 \\ x^3 \\ e^x \\ e^{-\frac{x^2}{\sigma^2}} \\ \vdots \end{bmatrix} \quad \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_1 x_2 \\ \sin x_1 x_2^2 \\ \vdots \end{bmatrix}$$

could even be ∞ dim

Feature space can get really large really quickly!

Higher order polynomials

$$\text{num. terms} = \binom{d+m-1}{d} = \frac{(d+m-1)!}{d!(m-1)!}$$



m – input features
d – degree of polynomial

even though dims of $\phi(x)$ are huge, fit model very ~~[illegible]~~ very quickly

grows fast!
d = 6, m = 100
about 1.6 billion terms