LOGISTIC REGRESSION

- CLASSIFICATION
- REGRESSION

$$\min_{w} \frac{1}{m} \sum_{i=1}^{m} (w^T x^i - y^i)^2$$

(*) $\mathbb{E}[Y|X] = w^T X$

EVEN IF (*) DOES NOT HOLD

CLASSIFICATION

WHAT IS LOSS FUNCTION FOR CLASSIFICATION?

OUR GUESS FOR LABEL $y^i \in \{0, +1\}$ IS GOING TO BE $\text{SIGN}(w^T x^i)$ FOR SOME VECTOR w.

- PENALIZED IF $\text{SIGN}(w^T x^i) \neq y^i$
  IF $\text{SIGN}(w^T x^i) = y^i$ (NO PENALTY)

EXAMINE QUANTITY $y^i \cdot (w^T x^i)$ IF QUANT IS POS NO PENALTY

" IS NEG PENALTY

0-1 LOSS $\varphi_{0-1}(z) = \begin{cases} 1 & \text{IF } z \leq 0 \\ 0 & \text{IF } z > 0 \end{cases}$

PE GD

SQ LOSS $(w^T x^i - y^i)^2$

CLASSIFICATION $\varphi_{0-1}(y^i \cdot w^T x^i)$

OPTIMIZATION PROBLEM ASSOCIATED WITH CLASSIFICATION

$$\min_w \frac{1}{m} \sum_{i=1}^{m} \ell_{0\text{-}1}\left(y^i \cdot w^T x^i\right)$$

WANT TO SOLVE.

WHEN DOES PERCEPTRON FIND A w WITH SMALL LOSS?.

RECALL THAT PERCEPTRON REQUIRED $\exists\, w$ s.t. $\forall x$

$$y \cdot w^T x > \rho \Rightarrow$$

CONVERGENCE # MISTAKES $< \frac{1}{\rho^2}$

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

QUESTION: WHAT IF THERE IS NO MARGIN?

THERE MIGHT NOT EXIST A w s.t. $\text{SIGN}(w^T x^i) = y^i \; \forall i$

OPTIMIZATION PROBLEM ASSOCIATED WITH CLASSIFICATION

$$\min_{w} \frac{1}{m} \sum_{i=1}^{m} \ell_{0\text{-}1}\left(y^i \cdot w^T x^i\right)$$

WHEN DOES PERCEPTRON FIND A w WITH SMALL LOSS?

WHAT IS THE COMPUTATIONAL COMPLEXITY OF THIS OPTIMIZATION PROBLEM?

BAD NEWS: THIS PROBLEM IS NP-HARD

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

QUESTION: WHAT IF THERE IS NO MARGIN?

THERE MIGHT NOT EXIST A w s.t. $\text{SIGN}(w^T x^i) = y^i \ \forall i$

RECALL THAT PERCEPTRON REQUIRED $\exists w$ s.t. $\forall x$

$$y \cdot w^T x > \rho \Rightarrow$$

CONVERGENCE # MISTAKES $< \frac{1}{\rho^2}$

OPTIMIZATION PROBLEM ASSOCIATED WITH CLASSIFICATION

$$\min_{w} \frac{1}{m} \sum_{i=1}^{m} \ell_{0\text{-}1}\left(y^i \cdot w^T x^i\right)$$

WHEN DOES PERCEPTRON FIND A w WITH SMALL LOSS?

WHAT IS THE COMPUTATIONAL COMPLEXITY OF THIS OPTIMIZATION PROBLEM?

BAD NEWS: THIS PROBLEM IS NP-HARD

"AGNOSTICALLY LEARNING A HALFSPACE"

AGNOSTIC LEARNING

QUESTION: WHAT IF THERE IS NO MARGIN?

THERE MIGHT NOT EXIST A w s.t. $\text{SIGN}(w^T x^i) = y^i \ \forall i$

RECALL THAT PERCEPTRON REQUIRED $\exists\, w$ s.t. $\forall x$

$$y \cdot w^T x > \rho \Rightarrow$$

CONVERGENCE # MISTAKES $< \frac{1}{\rho^2}$

TO SUMMARIZE

REGRESSION ⇝ CONVEX LOSS FUNCTION ✓

CLASSIFICATION ⇝ NON CONVEX LOSS (0-1) LOSS (BAD NEWS)

- IDEA LET'S RELAX THE 0-1 LOSS TO A DIFFERENT NICER LOSS
  "SURROGATE LOSSES" RELATED TO 0-1 LOSS BUT CONVEX

LET'S INTRODUCE A FEW LOSSES:

- $\psi_{\text{LOGISTIC LOSS}}$
- ψ_{HINGE}
- ψ_{EXP}

$$\ell_{\text{LOGISTIC}}(z) = \log\left(1 + e^{-z}\right)$$

$$\ell_{\text{LOGISTIC}}\left(y^i \cdot w^T x^i\right) = \log\left(1 + e^{-(y^i \cdot w^T x^i)}\right)$$

- IF $\underbrace{y^i w^T x^i}_{\text{MARGIN}} \ll 0 \Rightarrow \ell_{\text{LOGISTIC}}(y^i \cdot w^T x^i)$ IS LARGE

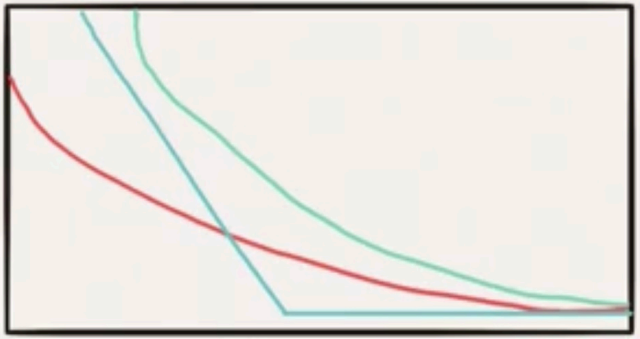
  $\gg 0 \Rightarrow \ell_{\text{LOGISTIC}}(y^i \cdot w^T x^i)$ IS SMALL (MOVING TO $\to 0$)

$$\ell_{\text{HINGE}}(z) = \max\{1 - z, 0\}$$

$\ell_{\text{HINGE}}(y^i \cdot w^T x^i)$ LARGE WHEN $y^i \cdot w^T x^i$ IS NEG

SMALL WHEN $y^i \cdot w^T x^i$ IS POS

$$\ell_{\text{EXP}}(z) = e^{-z}$$



$(z = y w^T x) \rightarrow$

LOGISTIC LOSS

HINGE LOSS

EXP LOSS

OPTIMIZATION PROBLEM ASSOCIATED WITH LOGISTIC LOSS?

$$L(w) = \frac{1}{m} \sum_{i=1}^{m} \log\left(1 + \exp\left(-y^i \cdot w^T x^i\right)\right)$$

$$\min_{w} L(w)$$

ENTER THE SIGMOID FUNCTION:

$$g(z) = \frac{1}{1 + e^{-z}}$$

As z GETS LARGE $g(z) \to 1$

z GETS SMALL $g(z) \to 0$

$$g(z) = \frac{1}{1+e^{-z}}$$

$Y \in \{0, +1\}$

$$\mathbb{E}[Y|X] = g(Y \cdot w^T X)$$

FOR SOME w $\exists w$

FACT: $g(z) + g(-z) = 1$

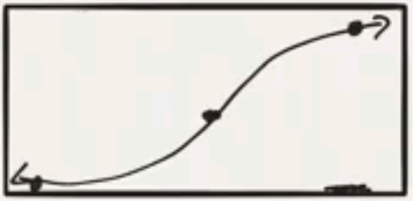
$$\frac{e^z}{e^z}\left(\frac{1}{1+e^{-z}}\right) + \frac{1}{1+e^z}$$

$$\Rightarrow \Pr[Y=1|X] = g(w^T X)$$

$$\frac{e^z}{1+e^z} + \frac{1}{1+e^z} = 1.$$

GIVEN X IF $w^T X$ IS LARGE

IF $w^T X$ IS NEG SMALL



$$Pr\left[y = y^i \,\middle|\, \underline{x}^i; \underline{w}\right] = g\left(y^i \cdot w^T x^i\right)$$

"MODEL FOR LOGISTIC REGRESSION"

$g(z) = \frac{1}{1+e^{-z}}$

GIVEN A TRAINING SET $S = \{(x^1, y^1), \ldots (x^m, y^m)\}$

WHAT IS THE MOST LIKELY w GIVEN THE TRAINING SET?

$$\text{LIKELIHOOD}(w) = \prod_{i=1}^{m} p(y = y^i \mid x^i; w) = \prod_{i=1}^{m} g\left(y^i \cdot w^T x^i\right)$$

MAX w

$$\text{LOG-LIKELIHOOD}(w) = = \sum_{i=1}^{m} \log g\left(y^i \cdot w^T x^i\right)$$

LOGISTIC LOSS

$$= -\sum_{i=1}^{m} \log\left(1 + \exp\left(-y^i \cdot w^T x^i\right)\right) = -m \cdot L(w)$$

NOW OUR GOAL WILL BE TO MINIMIZE LOGISTIC LOSS $L(w)$.

HOW SHOULD WE MINIMIZE LOGISTIC LOSS (w)?

IDEA: RUN GRADIENT DESCENT ON LOGISTIC LOSS.

THIS IS THE ALGORITHM FOR PERFORMING LOGISTIC REGRESSION!

LET'S SAY WE FIND w'; FOR FUTURE EXAMPLES WE LABEL THEM +1 WITH PROB $g(w'^T x)$

LET'S COMPUTE THE GRADIENT OF $L(w)$

LOGISTIC LOSS

$$\varphi_{\text{LOGISTIC}}(z) = \log(1+e^{-z})$$

THIS PRECISELY TELLS US HOW TO FIND MAX LIKELIHOOD w.

1. $$\varphi'_{\text{LOGISTIC}}(z) = \frac{-e^{-z}}{1+e^{-z}} = -\frac{1}{1+e^{z}} = -g(-z)$$

2. COMPUTE

$$\frac{\partial \varphi_{\text{LOGISTIC}}(y \cdot w^T x)}{\partial w_k} = -g(-y \cdot w^T x) \cdot y \cdot x_k$$

CHAIN RULE.

WITH THIS FORMULA WE CAN DIRECTLY APPLY GRADIENT DESCENT

WHAT HAPPENS IF WE HAVE MULTIPLE LABELS FOR y?

$y \in \{0, 1\}$

WHAT IF $y \in \{1, \dots, K\}$

MULTINOMIAL LOGISTIC REGRESSION: $w^1, \dots, w^{K-1}$

$$Pr[y = 1 \mid x] = \frac{e^{w^1 \cdot x}}{\sum_{i=1}^{K} e^{w^i \cdot x}}$$

$$Pr(y = j \mid x] = \frac{e^{w_j \cdot x}}{\sum_{j=1}^{K} e^{w}}$$

WHAT HAPPENS IF WE HAVE MULTIPLE LABELS FOR y? $y \in \{0,1\}$

WHAT IF $y \in \{1, \dots, K\}$

MULTINOMIAL LOGISTIC REGRESSION: $w^1, \dots, w^{K-1}$

$$Pr[y=1 \mid x] \propto e^{w^{1T} x}$$

$$Pr[y=K] = 1 - \sum_{i=1}^{K-1} Pr[y=i]$$

$$Pr[y=j \mid x] \propto e^{w^{jT} x}$$

WHAT IS THE ASSOCIATED LOSS?

CROSS-ENTROPY LOSS

GENERALIZATION OF LOGISTIC LOSS

((IMAGINE y IS A VECTOR OF LENGTH K WITH A 1 IN THE j^{th} POSITION IF CORRECT LABEL IS j)) (ONE-HOT ENCODING OF LABELS).

LET'S SAY OUR GUESS FOR THE PROB y HAS LABEL i IS p_i

$$-\sum_{i=1}^{K} y_i \log(p_i)$$

SOFTMAX ⇝ TURNS REAL-VALUED INTO PROBABILITIES

$W^T x$ VIA SIGMOID $W^T x \rightsquigarrow [0,1]$

INTO A PROBABILITY

$(z_1, \ldots, z_k) \rightsquigarrow \left(\frac{e^{z_1}}{Z}, \frac{e^{z_2}}{Z}, \ldots, \frac{e^{z_k}}{Z}\right)$

K COORDINATES

$$Z = \sum_{i=1}^{K} e^{z_i}$$