Assignment 1
Solutions

Machine Learning
CS 6375

Instructor:
Rishabh Iyer.

Extra (Regression)

Properties of Loss functions (Regression)

① $L(y, \hat{y}) = 0$ iff $y = \hat{y}$

② $L(y, \hat{y}) \geq 0 \quad \forall\, y, \hat{y}$

③ Good to have: Easy to minimize $L(y, \hat{y})$

Eg $L(y, \hat{y}) = \exp(y - \hat{y})$

satisfies ② but not ①

$\exp(y - \hat{y}) = 0$ if $y - \hat{y} = -\infty$

Assignment 1 solutions

Q 1] Which loss functions make sense for regression?

a) $L_i(w,b) = [f(w,b,x_i) - y_i]^3$

Does **not** make sense because this loss will be minimized when

$$f(w,b,x_i) - y_i \simeq -\infty$$

and not when $f(w,b,x_i) - y_i \simeq 0$

b) $L_i(w,b) = [f(w,b,x_i) - y_i]^4$

This **does** make sense because L_i is minimized when $f(w,b,x_i) \simeq y_i$

c) $L_i(w,b) = \exp[f(w,b,x_i) - y_i]$

Does **not** make sense because this will be minimized when

$$f(w,b,x_i) - y_i \simeq -\infty$$

$(\exp(-\infty) = 0)$

and not when $f(w,b,x_i) \simeq y_i$

d) $L_i(w,b) = \max(0, -y_i f(w,b,x_i))$

This loss fn also does not make sense because it is zero when $y_i f(w,b,x_i) \geq 0$ and not when $y_i \gtrsim f(w,b,x_i)$

For e.g. if $y_i = 2000$ & $f(w,b,x_i) = 1$ it will still be zero!

This is a Classification & not regression loss!

Note: Some students mentioned convexity. It is a good to have but not a requirement.

Gradients

$\nabla f_w(w,b,x_i) = x_i$, $\nabla f_b(w,b,x_i) = 1$.

a) $\nabla L_w^i(w,b) = 3[f(w,b,x_i) - y_i)^2 \nabla f_w(w,b,x_i)$

b) $\nabla L_w^i(w,b) = 4(f(w,b,x_i) - y_i)^3 \nabla f_w(w,b,x_i)$

c) $\nabla L_w^i(w,b) = \exp[f(w,b,x_i) - y_i] \nabla f_w(w,b,x_i)$

d) $\nabla L_w^i(w,b) = -1_{y_i f(w,b,x_i) < 0} y_i \nabla f_w(w,b,x_i)$

Gradients for b are same. Just replace w with b.

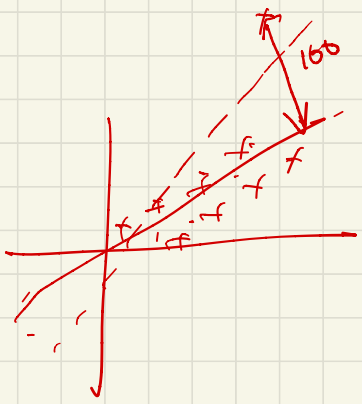
Bonus

$[f(w,b,x_i) - y_i]^4$ vs $[f(w,b,x_i) - y_i]^2$

The 4th power will be more sensitive to outliers. This can be a con for robust learning & a pro for fair learning (eg. imbalance)

Another bonus:

How will you fix a) & c) to make it into a valid loss fn?



100^4

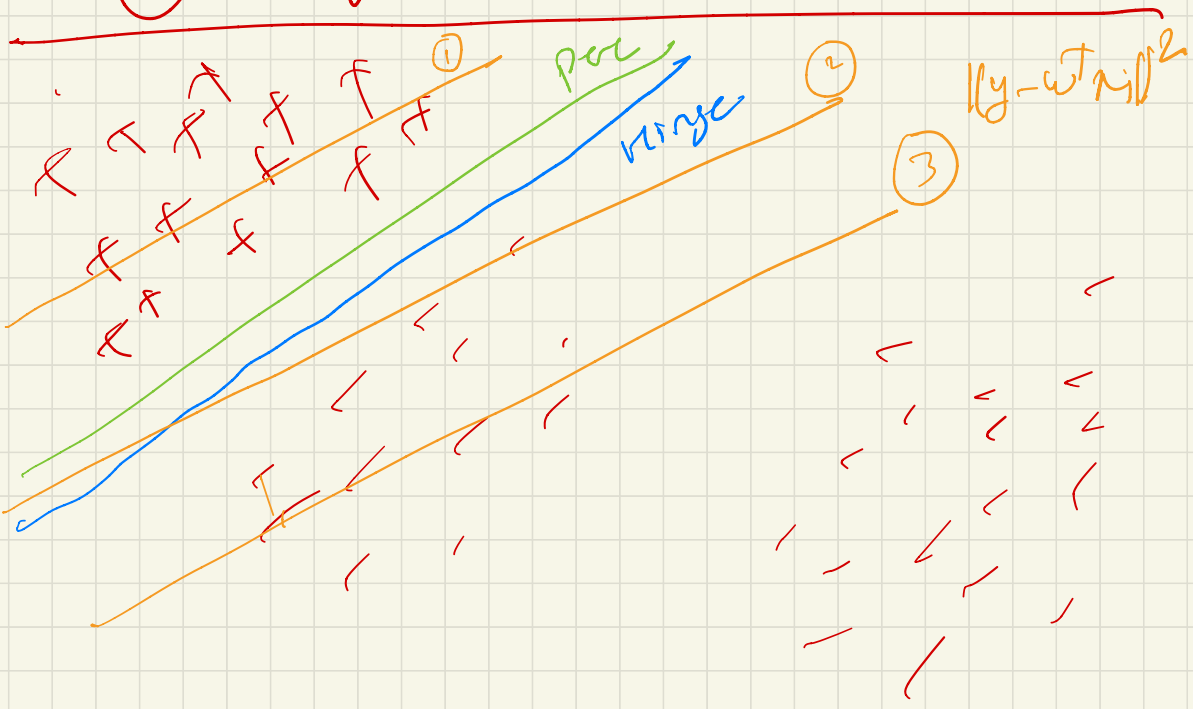
Extra (Classfn)

A loss fn $L(y, \hat{y})$ is good for Classfn if

Assume: $y \in \{0, 1\}$, $\hat{y} \in \mathbb{R}$

(1) $L(y, \hat{y}) = 0$ iff $y = \text{sign}(\hat{y})$

(2) $L(y, \hat{y}) \geq 0 \quad \forall\, y, \hat{y}$.



Q2]

a) $L_i(w,b) = \max(0, 1 - y_i f(w,b,x_i))^2$

This is squared Hinge loss It is a good choice for classification, because like the hinge loss

$L_i(w,b) = 0$ if $y_i f(w,b,x_i) \geq 1$

b) $L_i(w,b) = [y_i - f(w,b,x_i)]^4$

Note: this is a regression loss but is not good for classification because we want

$y_i \simeq \text{sign}(f(w,b,x_i))$

and not $y_i \simeq f(w,b,x_i)$

c) $L_i(w,b) = \exp(f(w,b,x_i) - y_i)$

This is neither good for classifn nor regression because

again, it is minimized when

$f(w,b,x_i) - y_i \simeq -\infty$

and not when

$y_i = \text{sign}(f(w,b,x_i)$

OR $y_i f(w,b,x_i) \gtrsim 0$

d) $L_i(w,b) = \exp(-y_i f(w,b,x_i))$

This is the exponential loss which is good for classification.

when $y_i f(w, b, x_i) \geq 0$,

the exp loss is ≤ 1

but is ≈ 0 only when

$y_i f(w, b, x_i) \simeq -\infty$

Gradients

a) $\nabla L_w^i(w, b, x_i) = 1_{y_i f(w,b,x_i) < 0} \, 2 y_i \nabla f_w(w, b, x_i)$

[same with b]

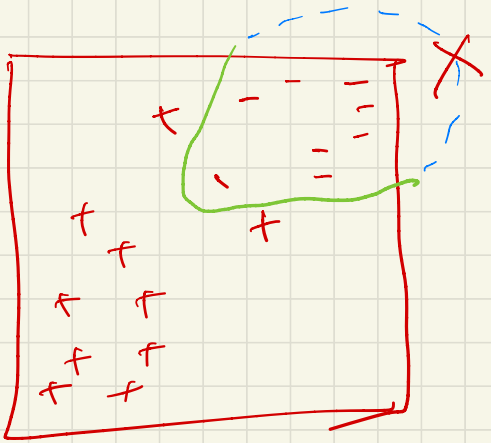
b) same as Q1.b)

c) same as Q1.c)

d) $\nabla L_w(w, b, x_i) = -\exp(-y_i f(w, b, x_i)) \, y_i \nabla f_w(w, b, x_i)$

[same with b]

Q3)



← Not ellipse/circle but parabola. ellipse/circle will overfit.

Features: $[1, x_1, x_2, x_1 x_2, x_1^2, x_2^2]$

Dataset is <u>linearly</u> sep with quadratic features.

Q4) (1) $\min_{w} \sum_{i=1}^{M} [w^T x^i + b - y^i]^2$

s.t $\|w - w_0\|^2 \leq k^2$

$$L(w, b, \lambda) = \sum_{i=1}^{M} [w^T x^i + b - y^i]^2 + \lambda(\|w - w_0\|^2 - k^2)$$

$$\nabla L_w = 2\sum_{i=1}^{M} (w^T x^i + b - y^i) x^i + 2\lambda(w - w_0) = 0$$

$$\nabla L_b = 2\sum_{i=1}^{M} (w^T x^i + b - y_i) = 0$$

$$\Rightarrow \quad b = \sum_{i=1}^{M} \frac{(y_i - w^T x^i)}{m}$$

Recall:

$$2\sum_{i=1}^{M}\left(w^T x^i + b - y^i\right)x^i + 2\lambda(w - w_0) = 0$$

sub. $b = \sum_{i=1}^{M}\frac{(y^i - w^T x^i)}{M}$

$$\Rightarrow 2\sum_{i=1}^{M}(w^T x^i - y^i)x^i + 2\sum_{i=1}^{M}\frac{(y^i - w^T x^i)}{M}\sum_{i=1}^{M}x^i + 2\lambda(w - w_0) = 0$$

Solve for $w = g(\lambda)$ and plug it back into dual.

$\beta \in \mathbb{R}^{n+1}$

Advanced approach

Write least squares as matrix operation.

$$L(\beta) = (y - X\beta)^T(y - X\beta) \qquad X = \begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}$$

where $\beta = [w, b]$

$$y = \begin{bmatrix} y_1 \\ y_M \end{bmatrix} \simeq \begin{bmatrix} x_1\beta \\ x_M\beta \end{bmatrix} = X\beta$$

Then, note:

$$\min_{\beta} \; (y - X\beta)^T(y - X\beta) \qquad \|y - X\beta\|^2$$

$$\text{s.t} \quad \|\beta - \beta_0\|^2 \le k^2$$

$$\Rightarrow L(\lambda, \beta) = (y - X\beta)^T(y - X\beta) + \lambda\left[\|\beta - \beta_0\| - k^2\right]$$

$$\frac{\partial L}{\partial \beta} = -2X^T(y - X\beta) + 2\lambda(\beta - \beta_0) = 0$$

$$\Rightarrow 2X^TX\beta - 2X^Ty + 2\lambda(\beta - \beta_0) = 0$$

$$\Rightarrow \beta(X^TX + \lambda I) = X^Ty + \lambda\beta_0$$

$$\Rightarrow \beta = (X^TX + \lambda I)^{-1}(X^Ty + \lambda\beta_0)$$

Substitute β into $L(\beta, \lambda)$ to get the dual

(+) Note: Connection to Closed form of Linear Regression?

$$L_R(\beta) = (y - X\beta)^T(y - X\beta)$$

$$\min_\beta L_R(\beta), \quad \nabla_\beta L_R(\beta) = 0$$

$$-2X^T(y - X\beta) = 0$$

$$X^TX\beta = X^Ty$$

$$\Rightarrow \beta = (X^TX)^{-1}X^Ty$$

Q4]

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

s.t $y_i(w^T x_i + b) \geq 1$

$\|w - w_0\|^2 \leq k^2$

$$L(w,b,\lambda,\mu) = \frac{1}{2}\|w\|^2 \quad \text{——— (1)}$$

$$+ \sum_{i=1}^{M} (1 - y_i(w^T x_i + b))\lambda_i + \mu[\|w - w_0\|^2 - k^2]$$

Now, $\nabla L_w = w - \sum_{i=1}^{M} \lambda_i y_i x_i + 2\mu_i(w - w_0) = 0$

$$\Rightarrow w[1 + 2\mu_i] = \sum_{i=1}^{M} \lambda_i y_i x_i + 2\mu_i w_0$$

$$\Rightarrow w = \frac{\sum_{i=1}^{M} \lambda_i y_i x_i + 2\mu_i w_0}{1 + 2\mu_i} \quad \text{——— (2)}$$

Sim, $\nabla L_b = -\sum_{i=1}^{M} \lambda_i y_i = 0$ [Constraint] — (3)

Substitute w (2) into (1) to get $g(\lambda)$ with constraints as (3)