- Linear classification with 0/1 loss
  - $S = \{(x_1, y_1), \dots (x_N, y_N)\}$, $x \in \mathbb{R}^d$, $y = \mathcal{Y} = \{0, 1\}$
  - $F = \{x \mapsto \mathbb{1}[\theta^T x + b \geq 0] : \theta \in \mathbb{R}^d, b \in \mathbb{R}\}$
  - The 0/1 loss is $\ell^{0,1}(\hat{y}, y) = \mathbb{1}[\hat{y} \neq y]$

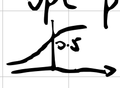
learning problem: $\arg\min \sum_{i=1}^{N} (\underbrace{\ell^{0,1}(f(x_i), y_i)}_{\mathbb{1}(f(x_i) \neq y_i)})$

error class 0 + err class 1

⚠ easy if classes are well separated
hard otherwise (NP-hard)

- logistic regression (replace the binary classification with a continuous function whose params we can opt on a convex opt pb)

sigmoid $\sigma(t) = \frac{1}{1+e^{-t}}$



$y = \{0, 1\}$

$\hat{y}_i = \theta^T x_i + b$ ← score for posit x_i

notation $\hat{\eta}_i = \sigma(\hat{y}_i) = \frac{1}{1+e^{-\hat{y}_i}}$ ← the class probab enties (probits)

$\hat{c}_i = \mathbb{1}[\hat{y}_i \geq 0] = \mathbb{1}[\hat{\eta}_i \geq \frac{1}{2}]$ ← predicted class

$\hat{y}_i$ is interpreted as $P[y = 1 | X = x_i; \theta, b]$

$(1 - \hat{y}_i)$ $P[y = 0 | \dots]$

Pb: learn θ, b

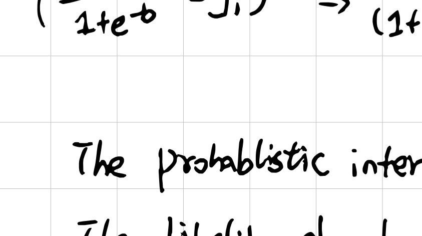
very naive idea: learn $\hat{\theta}, \hat{b}$: $\arg\min \mathbb{1}[\hat{c}_i \neq y_i]$

identical to previous one (hard)

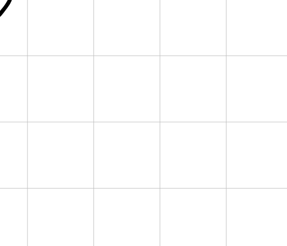
naive: $\arg\min \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$

the objective is continuous but not convex
⇒ hard to solve

Ex. show with a dataset with one example (0, 0) that the objective is non-convex.



→ obviously non-convex



$(\frac{1}{1+e^{-b}} - y_i)^2 \to \frac{1}{(1+e^{-b})^2}$ (sigmoid^2)

The probablistic interpretation:

The likelihood of mode θ, b:

$$\mathcal{L}(\theta, b) = \prod_{i=1}^{N} P(Y = y_i | X = x_i; \theta, b) = \prod_{\{i : y_i = 1\}} \hat{\eta}_i \cdot \prod_{\{i : y_i = 0\}} (1 - \hat{\eta}_i)$$

Neg log likelihood:

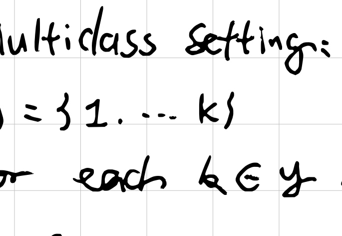
$$Nll = -\sum_{\{i : y_i = 1\}} \log \hat{\eta}_i - \sum_{\{i : y = 0\}} \log(1 - \hat{\eta}_i)$$

$$= \sum_{i=1}^{N} \underbrace{-y_i \ln \hat{\eta}_i - (1 - y_i) \ln(1 - \hat{\eta}_i)}_{\text{cross entropy } \ell^{CE}}$$

logistic regression pb.

$\hat{\theta}, \hat{b} = \arg\min \sum_{i=1}^{N} \ell^{CE}(\hat{\eta}_i, y_i)$

Ex. - draw $\ell^{CE}(\hat{\eta}, 1)$ and $\ell^{CE}(\hat{\eta}, 0)$



Multiclass setting:

- $\mathcal{Y} = \{1, \dots k\}$

for each $k \in \mathcal{Y}$, $\theta_k \in \mathbb{R}^d$, $b_k \in \mathbb{R}^d$

Score:

$(y_{i,1}, \dots y_{i,k}) = (\theta_1^T x_i + b_1 \dots \theta_k^T x_i + b_k)$

- predicted class: $\hat{c}_i = \arg\max_{k \in \{1, \dots k\}} \hat{y}_{ik}$

- $\text{softmax}\begin{pmatrix} t_1 \\ \vdots \\ t_k \end{pmatrix} = \frac{1}{\sum_{i=1}^{k} e^{t_k}} \begin{pmatrix} e^{t_1} \\ \vdots \\ e^{t_k} \end{pmatrix}$

Estimated class pro: $\hat{\eta}_i = (\hat{\eta}_{i,1} \dots \hat{\eta}_{i,k}) = \text{softmax}(\dots)$

- multiclass CE loss:

$$\ell^{CE}(\hat{\eta}_i, y_i) = \sum_{k=1}^{K} -\mathbb{1}_{[y_i = k]} \log \hat{\eta}_{i,k}$$

The logistic regression: the 2 views:

view 1: class probabilistic-estimation loss

$\arg\min \sum \ell^{CE}(\hat{\eta}_i, y_i)$

view 2: score loss $\arg\min \sum_i \underbrace{\ell^{\text{logistic}}(\hat{y}_i, y_i)}_{= \log(1 + e^{-y_i \hat{y}_i})}$

ex. develop view 1 $\xrightarrow{\text{identical}}$ view 2

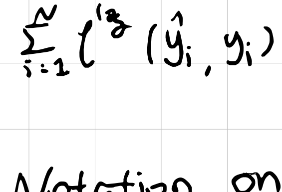
$\arg\min \sum \ell^{CE}(\hat{\eta}_i, y_i)$

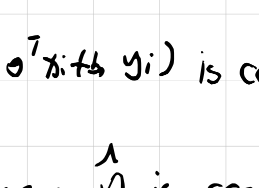
$= \dots \ell^{CE}(\frac{1}{1+e^{-\hat{y}_i}}, y_i)$

$= \dots \sum -y_i \log(\frac{1}{1+e^{-\hat{y}_i}}) - (1 - y_i) \log(1 - \frac{1}{1+e^{-\hat{y}_i}})$

$= \dots \sum \log(1 + e^{-y_i \hat{y}_i})$

$\log(1 + e^{-y\hat{y}})$ is convex in $\hat{y}$





the overall objective

$\sum_{i=1}^{N} \ell^{\log}(\hat{y}_i, y_i) = \sum_{i=1}^{N} \ell^{\log}(\theta^T x_i + b, y_i)$ is convex in θ and b

Notation on this page: $\hat{\eta}$ is score $\sigma(\hat{y})$

Neural nets : 2 views

1. outputs CPE

(cross prob estimation, CPE loss $\ell(\hat{\eta}, y)$)

2. Score $\hat{y}_i$ scoring loss (multiclass logistic loss) $\ell(\hat{y}, y)$

Proper CPE loss $\hat{\eta} \in [0,1]$ the risk of $\hat{\eta}(\cdot)$

$R^{\ell}(\hat{\eta}(\cdot)) = \mathbb{E}_{x,y \sim p}[\ell(\hat{\eta}(x), y)]$ with CPE loss ℓ

conditional risk : for $\eta \in [0,1]$ and $\hat{\eta} \in [0,1]$

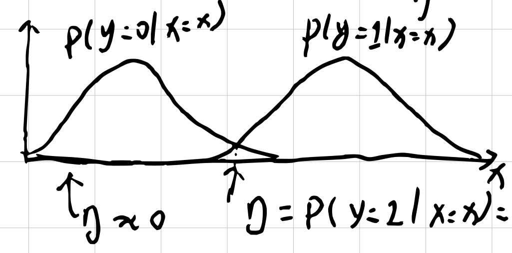
$C(\hat{\eta}, \eta) = \mathbb{E}_{y \sim Ber(\eta)}[\ell(\hat{\eta}, y)] = \eta \ell(\hat{\eta}, 1) + (1-\eta)\ell(\hat{\eta}, 0)$

note: $R^{\ell}(\hat{\eta}(\cdot)) = \mathbb{E}_x C(\hat{\eta}(x), \eta(x))$ where $\eta(x) = P(y=1|x)$

- A good CPE loss will be a proper loss:

def : a CPE loss is proper iff

$\forall \eta \in [0,1]$ $\eta \in \underset{\hat{\eta}}{\operatorname{argmin}}\ C(\hat{\eta}, \eta)$



$\mathcal{X} = \mathbb{R}$

$\mathcal{Y} = \{0,1\}$

$P:(y=1) = \frac{1}{2}$

$R^{\ell}(\hat{\eta}(\cdot)) = \mathbb{E}_{xy}[\ell(\hat{\eta}(x), y)]$

$= \mathbb{E}_x[\mathbb{E}_y[\ell(\hat{\eta}(x), y)|x]]$

$= \mathbb{E}_x[\eta(x) \times \ell(\hat{\eta}(x), 1) + (1-\eta(x)) \times \ell(\hat{\eta}(x), 0)]$

$\eta(x) = P(y=1|x)$

$= \mathbb{E}_x[C(\hat{\eta}(x), \eta(x))]$

Ex. show that cross entropy is proper

$C(\hat{\eta}, \eta) = \eta \ell(\hat{\eta}, 1) + (1-\eta)\ell(\hat{\eta}, 0)$

$= \eta(-\log \hat{\eta}) + (1-\eta)(-\log(1-\hat{\eta}))$

to show $\eta \in \underset{\hat{\eta}}{\operatorname{argmin}}\ C(\hat{\eta}, \eta)$

$\boxed{\text{def: a CPE loss is proper iff } \forall \eta \in [0,1]\ \eta \in \underset{\hat{\eta}}{\operatorname{argmin}}\ C(\hat{\eta}, \eta)}$

$\frac{\partial C}{\partial \hat{\eta}} = -\frac{\eta}{\hat{\eta}} + \frac{1-\eta}{1-\hat{\eta}}$

$= \frac{-\eta(1-\hat{\eta}) + \hat{\eta}(1-\eta)}{\hat{\eta}(1-\hat{\eta})} = 0$

$-\eta + \eta\hat{\eta} + \hat{\eta} - \hat{\eta}\eta = 0 \Rightarrow \hat{\eta} = \eta$

$\Rightarrow \eta \in \underset{\hat{\eta}}{\operatorname{argmin}}\ C(\hat{\eta}, \eta)$

ℓ^{CE} is (strictly) proper

$\inf_f R^{\ell}(f) = \mathbb{E}_x \inf_{\hat{\eta}} C(\hat{\eta}, \eta(x)) = \mathbb{E}_x C(\eta(x), \eta(x))$

cross entropy: $\mathbb{E}_x C(\eta(x), \eta(x)) = \mathbb{E}_x H(\eta(x))$ where $H(\cdot)$ is entropy.

- $\ell(\hat{\eta}, y)$ → in $\{0,1\}$, could also be intepreted as a parameter

param of a Bernoulli distri

ex. $\ell(10\%, 0)$

Ber(10%) Ber(0%)

Kullback-Leibler divergence : compare discrete distribution

two discrete distributions p and q on $\mathcal{Y}$

KL - divergence

$KL(P||q) = \sum_{y \in \mathcal{Y}} p(y) \log \frac{p(y)}{q(y)}$

property :

$KL(P||q) = 0$ if $p = q$

> 0 if $p \neq q$

$KL(P||q) \neq KL(q,p)$

Compare $Ber(\hat{\eta})$ with $Ber(y)$

$KL(Ber(y), Ber(\hat{\eta})) = y \times \log \frac{y}{\hat{\eta}} + (1-y)\log \frac{1-y}{1-\hat{\eta}}$

$= \underbrace{-y\log\hat{\eta} - (1-y)\log(1-\hat{\eta})}_{\ell^{CE}(\hat{\eta}, y)}$

$+ \underbrace{y\log y + (1-y)\cdot\log(1-y)}_{-\text{Entropy}(Ber(y)) = 0}$

Assume $0 \log 0 = 0$

$\ell^{CE}(\hat{\eta}, y) = KL(Ber(y), Ber(\hat{\eta}))$

Scoring loss

formulated: $y = \{-1, 1\}$ $\quad \ell(\hat{y}, y)$ logistic losss

$\ell(\hat{y}, y) = \phi(\hat{y} y)$ $\quad \phi$ - loss or margin loss

$\ell^{0,1}(\hat{y}, y) = \mathbb{1}[y\hat{y} < 0]$ $\quad \phi(\cdot) = \mathbb{1}[\cdot < 0]$

$\ell^{\text{logistic}}(\hat{y}, y) = \log(1 + e^{-\underbrace{y\hat{y}}_{\text{margin}}})$ $\quad \phi(z) = \log(1 + e^{-z})$

$\ell^{\text{hinge}} = \max(1 - y\hat{y}, 0)$

The risk of scoring loss:

$\eta(x) = P(y = 1 \mid X = x)$

$\hat{y} \in \mathbb{R} \cup \{-\infty, \infty\}$ is a score any $\hat{y}(\cdot)$: $X \mapsto \mathbb{R} \cup \{-\infty, \infty\}$ is a scoring function.

def: for any $\hat{y} \in \mathbb{R}$, the conditional risk for scoring loss ℓ $\quad C(\hat{y}, \eta) = \mathbb{E}_{y \sim \{1 \text{ w.p. } \eta,\ -1 \text{ w. } 1-\eta\}} [\ell(\hat{y}, y)]$

$$= \eta \ell(\hat{y}, 1) + (1-\eta)\ell(\hat{y}, -1)$$

def: the risk for scoring func $\hat{y}(\cdot)$: $\mathbb{R}^d \to \mathbb{R}$ for loss ℓ is

$$R^{\ell}(\hat{y}(\cdot)) = \mathbb{E}_{x,y \sim p}[\ell(\hat{y}(x), y)] = \mathbb{E}_x[C(\hat{y}(x), \eta(x))]$$

Calibration of scoring losses:

what's "good" scoring loss?

def: A loss ℓ is calibrational if

if $\eta \in [0, \frac{1}{2}[$ then $\quad \inf_{\hat{y} < 0} C(\hat{y}, \eta) < \inf_{\hat{y} \geq 0} C(\hat{y}, \eta)$

if $\eta \in]\frac{1}{2}, 1]$ then $\quad \inf_{\hat{y} > 0} C(\hat{y}, \eta) < \inf_{\hat{y} \leq 0} C(\hat{y}, \eta)$

Thm: if ℓ is calibrated then

if $\hat{y}(\cdot)$ is the measurable function minimizing R^{ℓ},

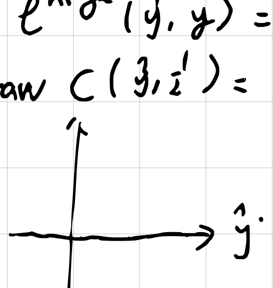
then it also minimizing $R^{0/1}$

Intuition: a measurable fun learnt to min calibrational loss will also min 0/1 lo

$C^{(\hat{y}, \eta)} = \eta \ell(\hat{y}, 1) + (1-\eta)\ell(\hat{y}, -1)$

Ex. $\ell^{\text{hinge}}(\hat{y}, y) = \max(0, 1 - \hat{y} y)$

draw $C(\hat{y}, \frac{1}{2}) = \eta(1 - \hat{y}) + (1-\eta)(1 + \hat{y}) =$



Thm: Any convex ϕ-loss with $\phi'(0) < 0$ is well-calibrated.

$\Rightarrow$ all scoring function we saw are - - -