Learning in the Classification Setting
From the 0/1 loss to Convex Surrogate Losses

1) Linear Classification
   - Learning with the 0/1 loss
   - A surrogate learning problem: Logistic Regression
   - The two views of logistic regression

2) CPE-Losses (Class Probability Estimate Losses)
   - Definition of CPE-losses
   - Proper losses
   - the Cross-Entropy Loss

3) Scoring Losses
   - Definition of Scoring Losses
   - Well Calibrated Losses

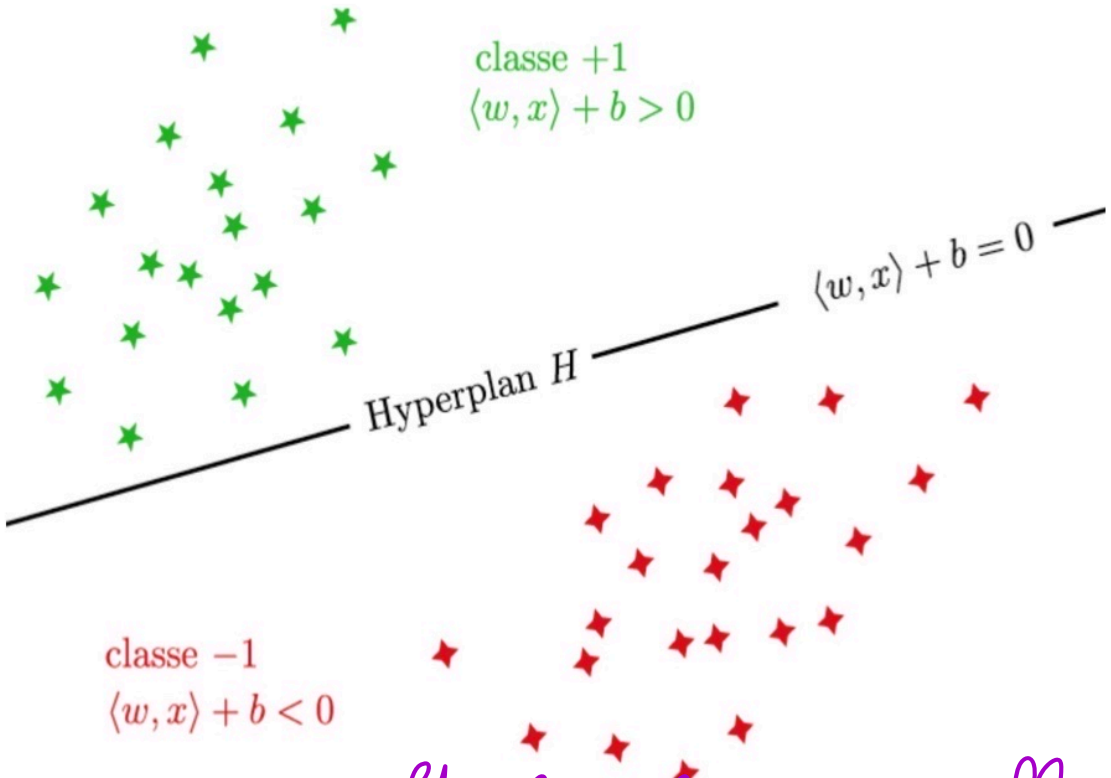
1) Linear classification

linear classification with the 0/1 loss

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

notation:

$$\theta^T x = \langle \theta, x \rangle = \sum_{j=1}^{d} \theta^{(j)} x^{(j)}$$



learning problem:

$$\underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{N} \underbrace{\ell^{0-1}(f(x_i), y_i)}_{\mathbb{1}[f(x_i) \neq y_i]}$$

$$\Updownarrow$$

$$\underset{\theta, b}{\operatorname{argmin}} \sum_{i : y_i = 1} \mathbb{1}[\theta^T x_i + b < 0] + \sum_{i : y_i = 0} \mathbb{1}[\theta^T x_i + b \geq 0]$$

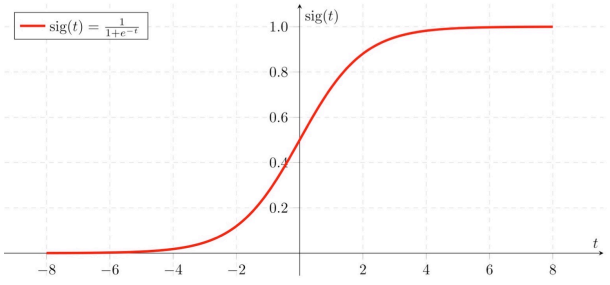
⚠ easy pb iff the classes are well separated
Hard (NP-hard) otherwise

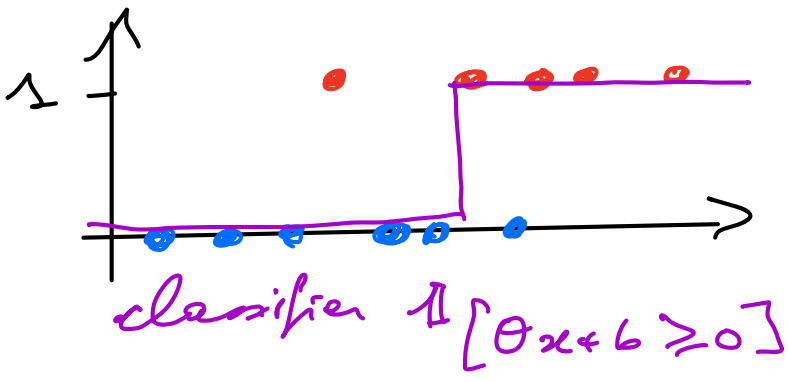
Linear classification with the logistic regression framework

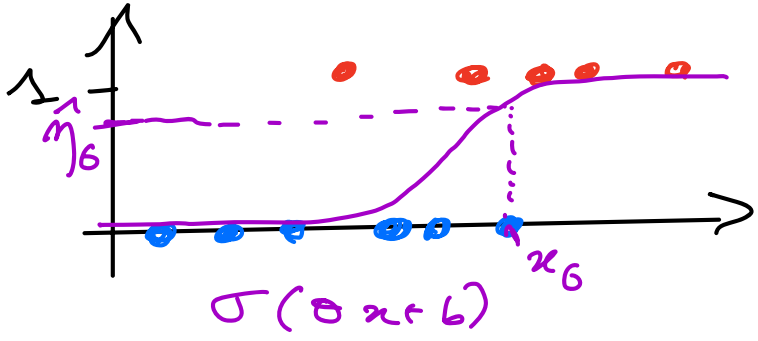
- Idea: replace the binary classifier with a continuous function whose parameters we can optimize as a convex opt pb.

- The sigmoid function

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







- why the sigmoid function?

Linear classification with logistic regression : which loss ?

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

Notations: (dependency on θ, b, omitted for brievety)

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

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

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

- $\hat{\eta}_i$ is interpreted as $P[Y=1 \mid X=x_i ; \theta, b]$
- $(1-\hat{\eta}_i)$ " " " $P[Y=0 \mid X=x_i ; \theta, b]$
- Pb : learn θ, b

Very naïve Idea: learn $\hat{\theta}, \hat{b} = \text{argmin} \sum_{i=1}^{N} \mathbb{1}[\hat{c}_i \neq y_i]$
Naive because identical to previous problem, so hard

Naïve Idea : learn $\hat{\theta}, \hat{b} = \text{argmin} \sum_{i=1}^{N} (\hat{\eta}_i - y_i)^2$
Naïve because the objective is continuous, but not convex ⇒ hard to solve.

Exercise

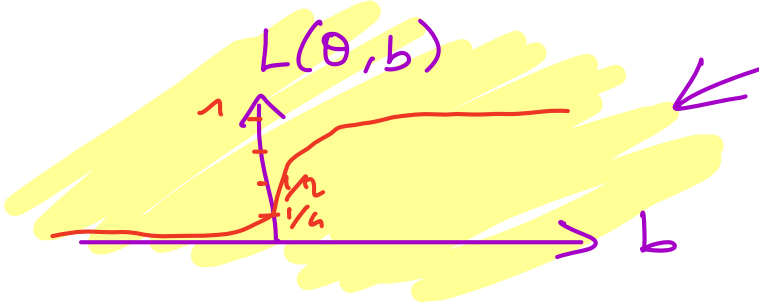
show with a dataset of one example $(0, 0)$ that the objective is not convex.

$$\text{argmin} \underbrace{\sum_{i=1}^{N} (\hat{\eta}_i - \eta_i)^2}_{L(\theta, b)}$$

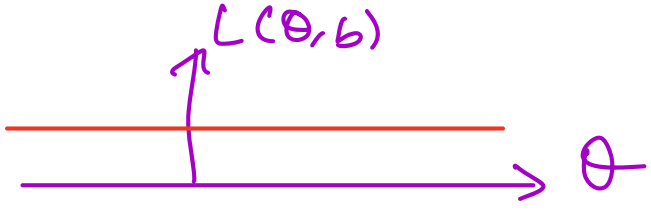
recall:

$$\hat{y}_i = \theta^T x_i + b$$

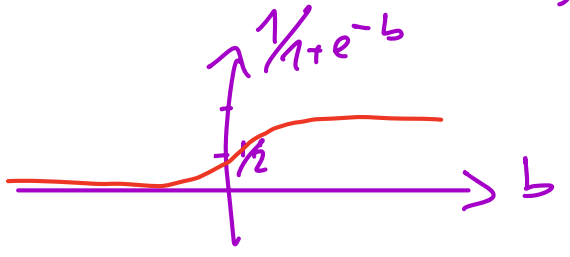
$$\hat{\eta}_i = \sigma(\hat{y}_i) = \frac{1}{1 + e^{-\hat{y}_i}}$$



obviously non convex



Here, $L(\theta, b) = \left(\frac{1}{1 + e^{-b}}\right)^2$



Linear classification with logistic regression : which loss ?

- We will use the probabilistic interpretation of logistic regression to get a good loss.

- The likelihood of the model θ, b is

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

- Making the data the most probable w.r.t. θ, b amounts to maximize $\mathcal{L}(\theta, b)$ or to minimize the negative log likelihood $NLL(\theta, b)$

$$NLL(\theta, b) = -\log \mathcal{L}(\theta, b)$$

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

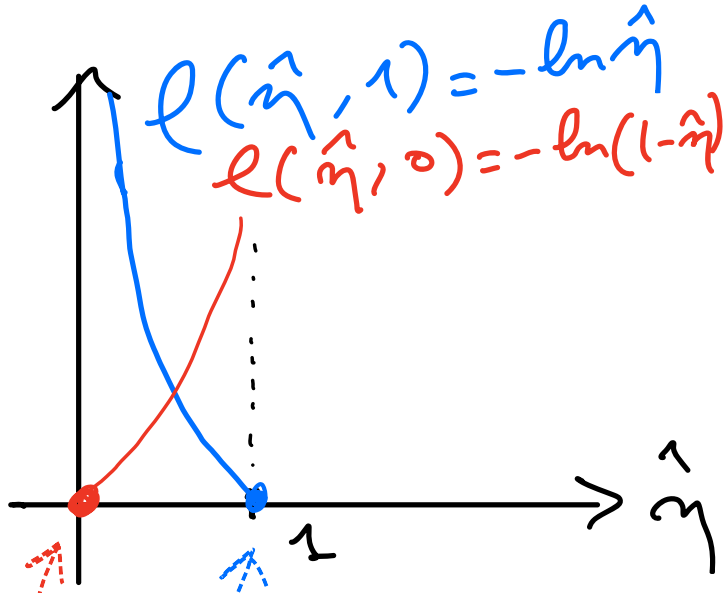
$$= \sum_{i=0}^{N} \underbrace{- y_i \ln \hat{\eta}_i - (1 - y_i) \ln (1 - \hat{\eta}_i)}_{\ell^{CE}(\hat{\eta}_i, y_i) = \text{cross-entropy loss}}$$

Logistic regression pb

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

Exercise

- draw $\ell^{CE}(\hat{\eta}, 1)$ and $\ell^{CE}(\hat{\eta}, 0)$
- Based on the drawing, show $\underset{\hat{\eta}}{\operatorname{argmin}}\, \ell^{CE}(\hat{\eta}, 0) = 0$ and $\underset{\hat{\eta}}{\operatorname{argmin}}\, \ell^{CE}(\hat{\eta}, 1) = 1$

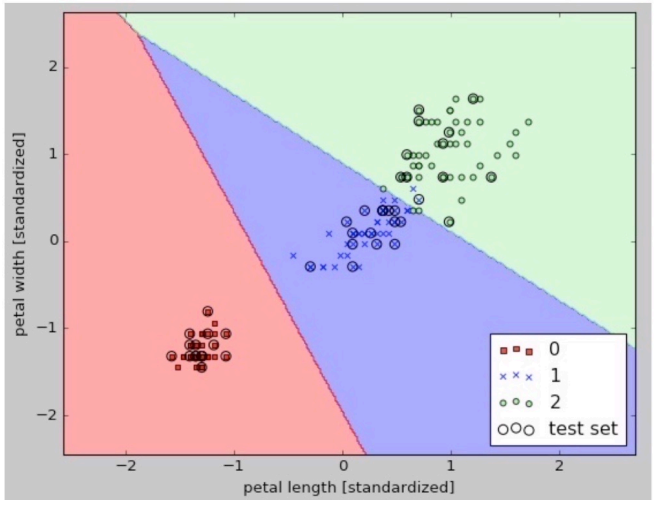


logistic regression : the multiclass setting

- $\mathcal{Y} = \{1..K\}$
- for each $k \in \mathcal{Y}$, we have $\theta_k \in \mathbb{R}^d$, $b_k \in \mathbb{R}^d$
- for each example x_i, the scores of x_i for each class is

$$(\hat{y}_{i,1} \cdots \hat{y}_{i,K}) = (\theta_1^\top x_i + b_1, \dots, \theta_K^\top x_i + b_K)$$

- predicted class: $\hat{c}_i = \underset{k \in \{1..K\}}{\operatorname{argmax}} \hat{y}_{ik}$



- replace sigmoid by softmax:

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

- Estimated class probability $\hat{\eta}_i = (\hat{\eta}_{i,1} \cdots \hat{\eta}_{i,K}) = \text{softmax}(\hat{y}_{i,1} \cdots \hat{y}_{i,K})$
- multiclass cross entropy loss

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

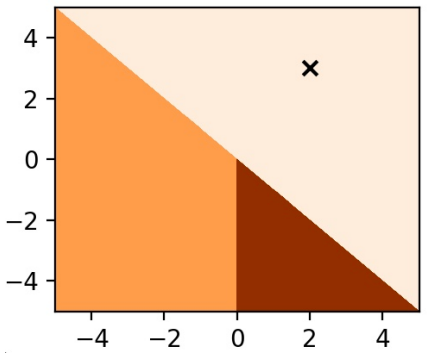
logistic regression : the multiclass setting, example

three classes : $\mathcal{Y} = \{dog, cat, mouse\}$

$x = (2, 3)$

parameters

$\theta_{dog} = (1, 1)^T$ $b_{dog} = 0$	$\theta_{cat} = (-1, 0)^T$ $b_{cat} = 0$	$\theta_{mouse} = (1, -1)^T$ $b_{mouse} = 0$
scores $\hat{y}_{dog} = 5$	$\hat{y}_{cat} = -2$	$\hat{y}_{mouse} = -1$
CPEs $\exp(\hat{y}_{dog}) = 148$ $\sum_c \exp(\hat{y}_c) = 149$ $\hat{\eta}_{dog} = 99,6\%$	$\exp(\hat{y}_{cat}) = 0,1$ $\hat{\eta}_{cat} = 0,1\%$	$\exp(\hat{y}_{mouse}) = 0,4$ $\hat{\eta}_{mouse} = 0,2\%$
predicted class $\hat{c} = dog$		



The logistic Regression: the two views

(in the bi-class setting)

View 1: Working with cross-entropy loss $\ell^{CE}(\hat{\eta}, \eta)$ with $\eta \in \{0,1\}$ which is a CPE-loss (class probability estimate)

$$\hat{\theta}, \hat{b} = \arg\min_{\theta, b} \sum_i \ell^{CE}(\hat{\eta}_i, \eta_i)$$

View 2: work with the logistic loss on scores with $y \in \{-1,1\}$

The logistic loss $\ell^{logistic}(\hat{y}, y)$ is a scoring loss

$$\hat{\theta}, \hat{b} = \arg\min_{\theta, b} \sum_i \ell^{logistic}(\hat{y}_i, y_i)$$

with $\ell^{logistic}(\hat{y}_i, y_i) = \ln(1 + e^{-y_i \hat{y}_i})$

Both views are identical, and are convex in θ, b

Exercise: develop view 1 to show it is identical to view 2.

The logistic Regression: the two views

Exercise: Develop view 1 to show it is identical to view 2.

Solution:

The logistic Regression: the two views

Exercise: develop view 1 to show it is identical to view 2.

solution:

$$\ell^{CE}(\hat{\eta}, 1) = -\ln \hat{\eta} = -\ln \frac{1}{1+e^{-\hat{y}}} = \ln\left(1+e^{-\hat{y}}\right)$$

$$\ell^{CE}(\hat{\eta}, 0) = -\ln(1-\hat{\eta}) = -\ln\left(1 - \frac{1}{1+e^{-\hat{y}}}\right) = -\ln \frac{e^{-\hat{y}}}{1+e^{-\hat{y}}}$$

$$= -\ln\left(\frac{1}{1+e^{\hat{y}}}\right) = \ln\left(1+e^{\hat{y}}\right)$$

$$\ell^{CE}(\hat{\eta}, y) = \ln\left(1 + e^{-y\hat{y}}\right)$$

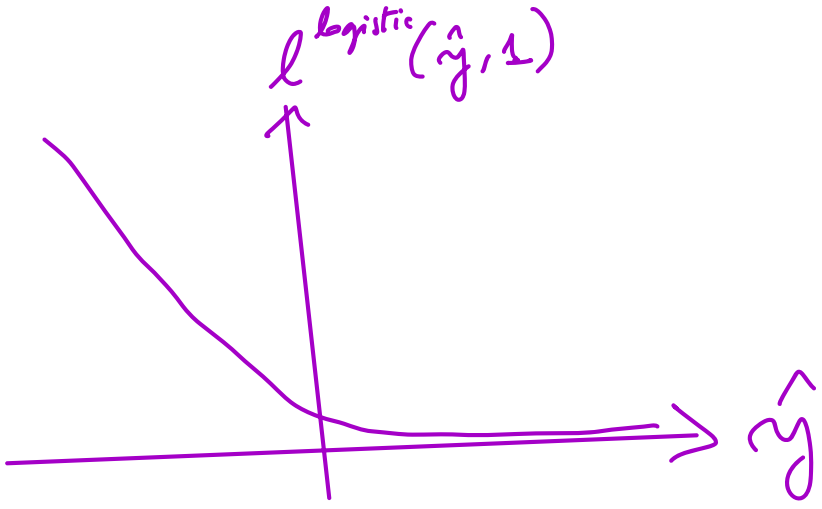
$y \in \{0,1\}$ — cross entropy loss

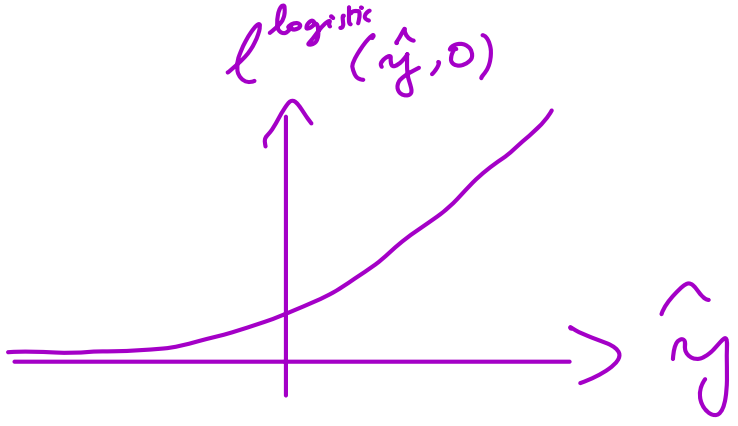
$y \in \{-1,1\}$ — logistic loss

Convexity of the learning problem

$\ell^{\text{logistic}}(\hat{y}, y) = \ln(1 + e^{-y\hat{y}})$ is convex in $\hat{y}$

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





the overall objective

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

is convex in θ and b

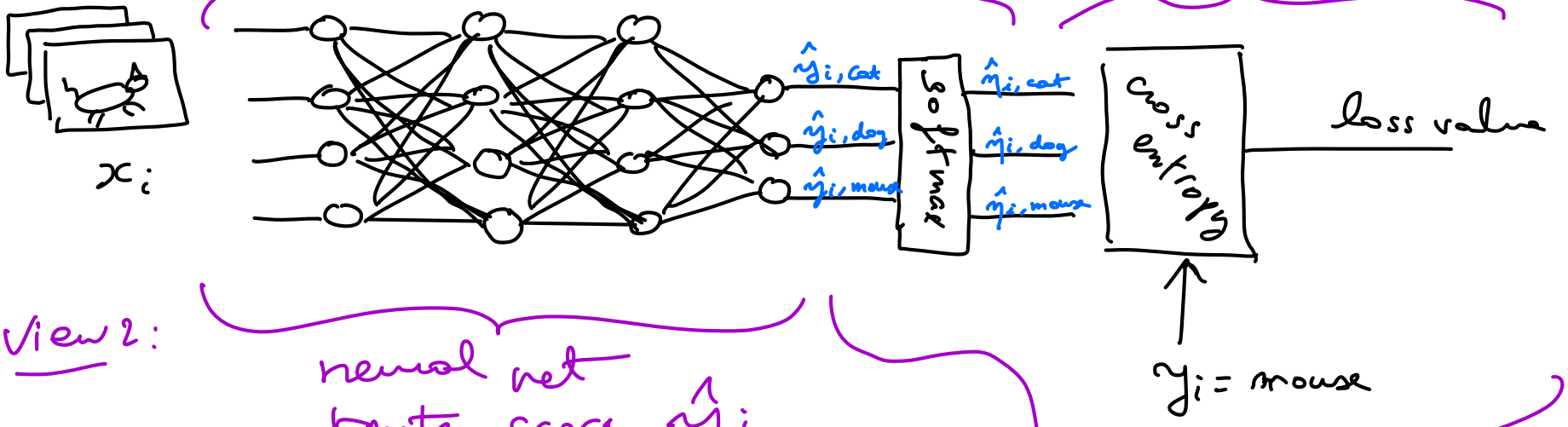
From logistic Regression to neural nets: the two views

- We train a neural net on 64x64 images with $y = \{\text{cat}, \text{dog}, \text{mouse}\}$

$$x = \mathbb{R}^{64 \times 64}$$

View 1: Neural net outputs CPE ($\hat{\eta}_i$) (class proba estimates)

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



View 2: neural net outputs scores $\hat{y}_i$

Scoring loss (multiclass logistic loss) $\ell(\hat{y}, y)$

2) CPE losses

Class Probability Estimated

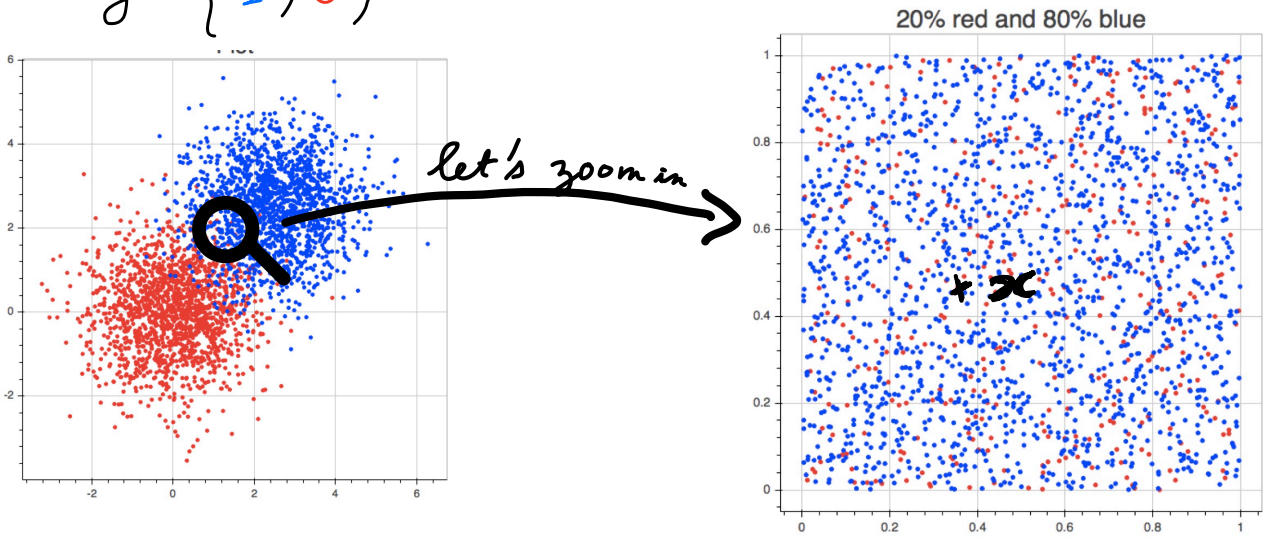
2) CPE-losses (for $\mathcal{Y} = \{0,1\}$)

- Recall that a CPE loss is of the form

  $\ell(\hat{\eta}, \eta)$ (e.g. cross entropy).

- What are good CPE-losses?

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



- The point x is in a region where $\eta = P(Y=1|X) = 80\%$
- So a neural network trained with a "good" CPE loss should predict the correct probability $\hat{\eta} = 80\%$
- The average loss incurred by points in this region will be $\eta\,\ell(\hat{\eta}, 1) + (1-\eta)\,\ell(\hat{\eta}, 0)$

3) Risk of CPE-losses

Notation:

- $\hat{\eta} \in [0,1]$ is a probability
- $\hat{\eta}(\cdot) \in \mathbb{R}^d \mapsto [0,1]$ is a CPE prediction fction (eg. a neural net with softmax)

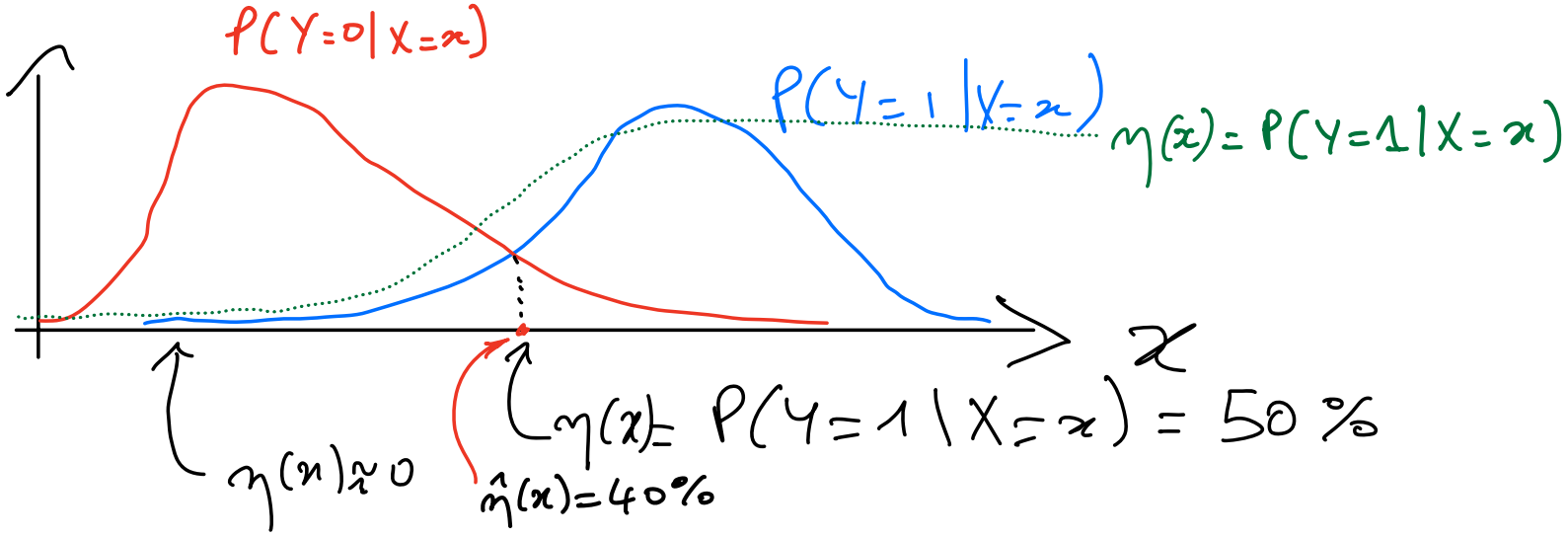
- Let us first adapt our definition of the risk to new losses

def: the risk of a prediction function $\hat{\eta}(\cdot)$ with CPE-loss ℓ is

$$R^{\ell}(\hat{\eta}(\cdot)) = \mathbb{E}_{X,Y \sim P}[\ell(\hat{\eta}(X), Y)]$$

- To capture the intuition of the previous slide, we need the conditional risk

4) conditional risk of CPE losses



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

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

$$\begin{aligned} R^{\ell}(\hat{\eta}(\cdot)) &= \mathbb{E}_{X,Y \sim P}[\ell(\hat{\eta}(X), Y)] \\ &= \mathbb{E}_X\left[\mathbb{E}_Y[\ell(\hat{\eta}(X), Y) \mid X]\right] \\ &= \mathbb{E}_X\left[\eta(X) \times \ell(\hat{\eta}(X), 1) + (1-\eta(X)) \times \ell(\hat{\eta}(X), 0)\right] \\ &= \mathbb{E}_X\left[C(\hat{\eta}(X), \eta(X))\right] \end{aligned}$$

5) Proper CPE losses

- A "good" CPE loss will be a proper loss

def: a CPE loss is proper iff

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

Exercises

1)- show the cross entropy is proper

Let us show that $\eta = \underset{\hat{\eta}}{\operatorname{argmin}}\, C(\hat{\eta}, \eta)$

$$C(\hat{\eta}, \eta) = -\eta \ln \hat{\eta} - (1-\eta)\ln(1-\hat{\eta})$$

$$\frac{d}{d\hat{\eta}} C = -\frac{\eta}{\hat{\eta}} + \frac{(1-\eta)}{1-\hat{\eta}} \triangleq 0 \Rightarrow \frac{\eta}{\hat{\eta}} = \frac{1-\eta}{1-\hat{\eta}}$$

$$\Rightarrow \eta(1-\hat{\eta}) = \hat{\eta}(1-\eta) \Rightarrow \eta = \hat{\eta} \Rightarrow \ell^{CE} \text{ is (strictly) proper}$$

2)- show (latter) that the MSE mean square error is strictly proper.

3)- Write the Bayes risk $(= \min_f R^\ell(f))$ for the case ℓ is proper. Apply to cross entropy

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

cross entropy, $\mathbb{E}_X\, C(\eta(X), \eta(X)) = \mathbb{E}_X\, H(\eta(X))$ where $H(\cdot)$ is entropy

6) More on CPE-losses and the cross-entropy

- $\ell(\hat{\eta}, y)$ ← in $[0,1]$, could also be interpreted as a parameter of Bernoulli:
  - $\hat{\eta}$: parameter of a Bernoulli distribution

ex: $\ell(10\%, 0)$ can be interpreted as a dissimilarity between two distributions $\mathrm{Ber}(10\%)$ and $\mathrm{Ber}(0\%)$

- There is a standard way to compare discrete distributions: Kullback-Leibler divergence

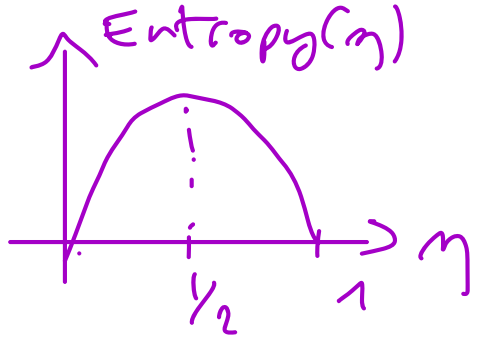
for two discrete distributions p and q on Y,

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

- property: $KL(p \| p) = 0$, $KL(p \| q) > 0$ if $p \neq q$

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

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



I compare $Ber(\hat{\eta})$ with $Ber(y)$

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

$$= \underbrace{-y \ln \hat{\eta} - (1-y) \ln (1-\hat{\eta})}_{\ell^{CE}(\hat{\eta}, y)} + \underbrace{y \ln y + (1-y) \ln (1-y)}_{-\text{Entropy}(Ber(y)) = 0}$$

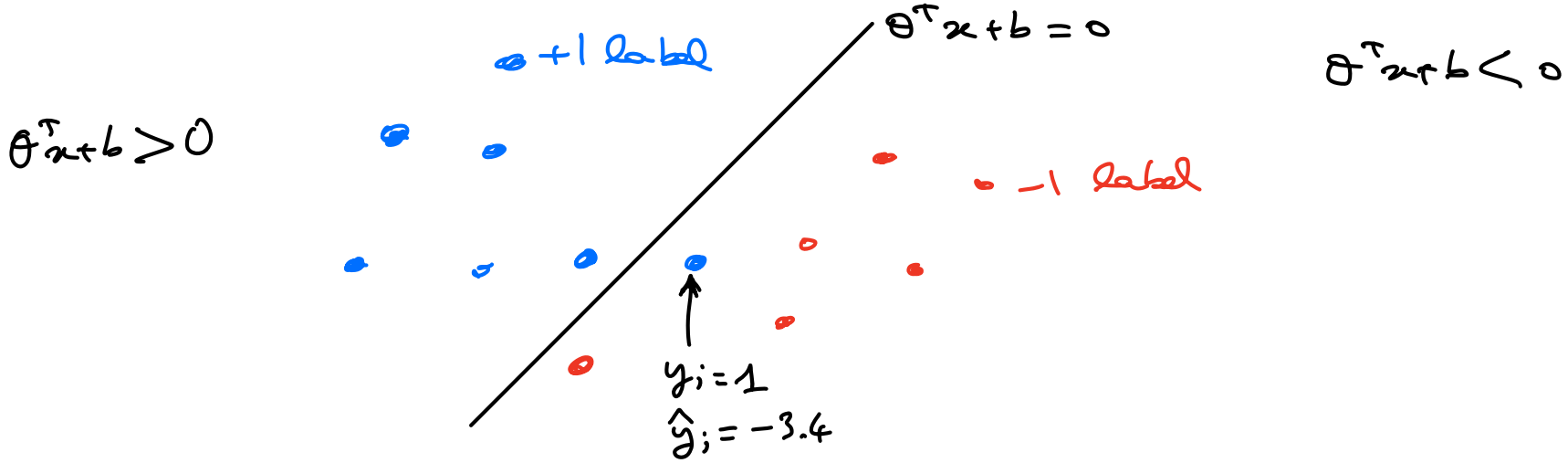
(Assumption $0 \ln 0 = 0$)

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

3. Scoring Losses

3) Scoring losses

- $y = \{-1, 1\}$
- $\ell(\hat{y}, y)$ ex: logistic loss
- with linear classifier $\hat{y}_i = \theta^T x + b$



- Most scoring losses can be formulated this way: $\ell(\hat{y}, y) = \phi(\hat{y} y)$ for some function ϕ
  These are called ϕ-losses or margin losses
- examples of ϕ-losses

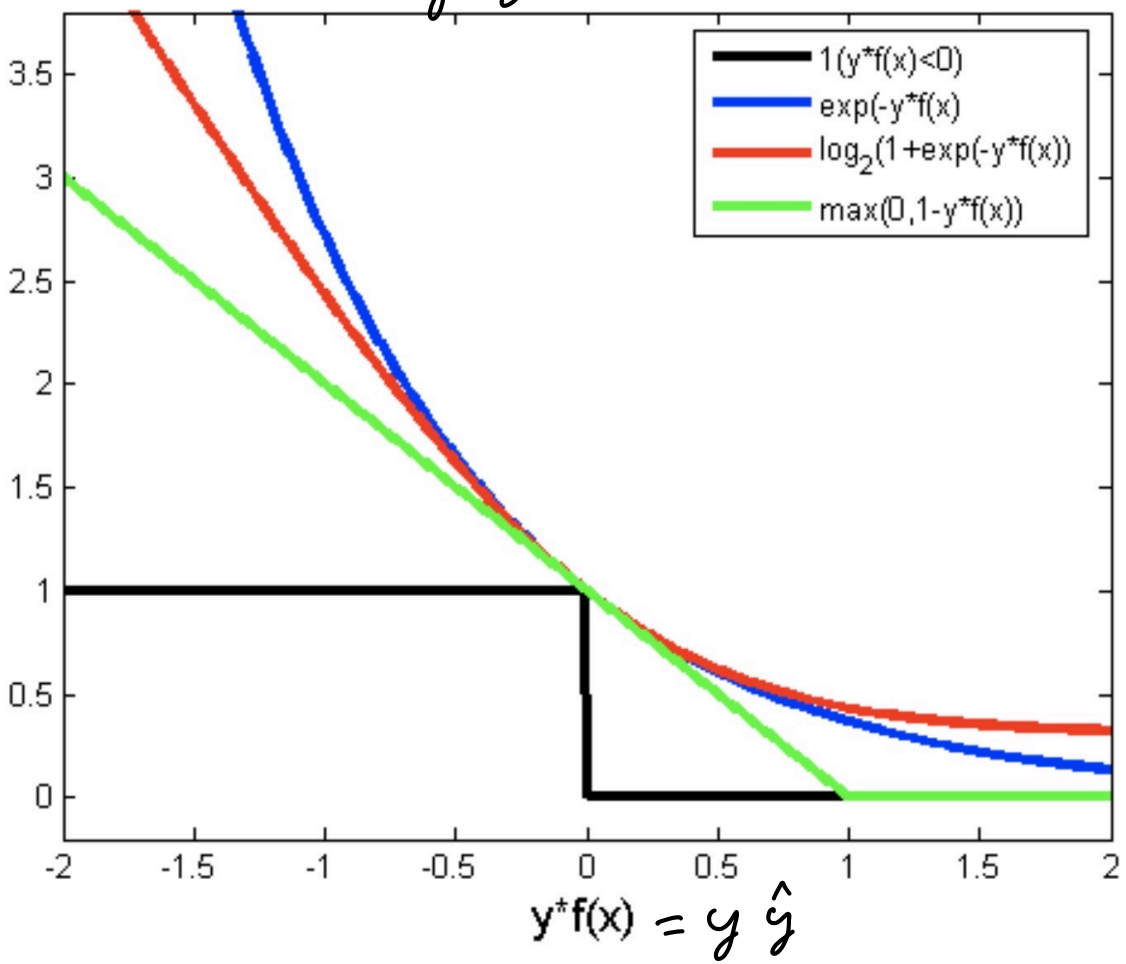
Scoring losses

$\ell^{0/1}(\hat{y}, y) = \mathbb{1}[y\hat{y} < 0]$ is a ϕ-loss with $\phi(\cdot) = \mathbb{1}[\cdot < 0]$

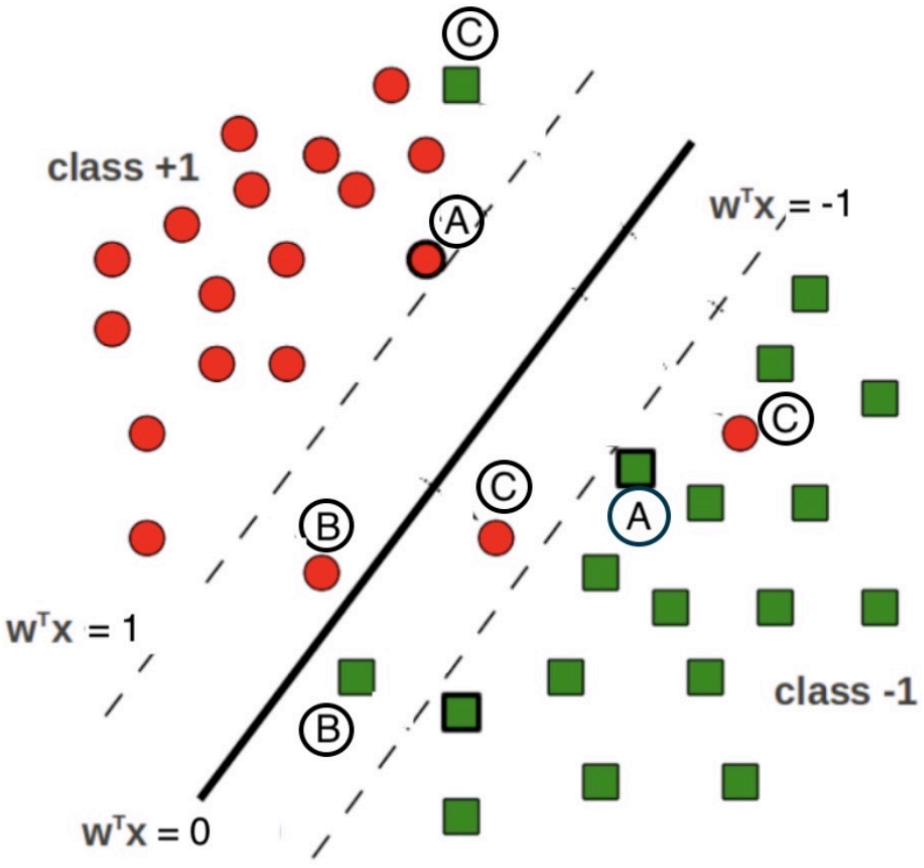
(margin)

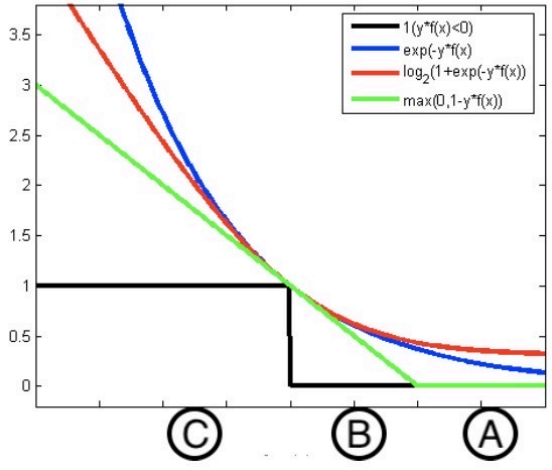
$\ell^{\text{logistic}}(\hat{y}, y) = \ln(1 + e^{-y\hat{y}})$ is a ϕ-loss with $\phi(z) = \ln(1 + e^{-z})$

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



Scoring losses





The risk of a scoring loss

We note $\eta(x) = P(Y=1 \mid X=x)$

$\hat{y} \in \mathbb{R} \cup \{-\infty, \infty\}$ is a score and $\hat{y}(\cdot) \in \mathcal{X} \mapsto \mathbb{R} \cup \{-\infty, \infty\}$ is a scoring function (e.g. a neural net)

def: For any $\hat{y} \in \mathbb{R}$, the conditional risk for scoring loss ℓ is

$$C(\hat{y}, \eta) = \mathbb{E}_{Y \sim \begin{cases} 1 \text{ w.p. } \eta \\ -1 \text{ w.p. } 1-\eta \end{cases}} [\ell(\hat{y}, Y)]$$

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

def: The risk of scoring function $\hat{y}(\cdot): \mathbb{R}^d \to \mathbb{R}$ 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

Same as before: what are "good" scoring losses?

def: A loss ℓ is calibrated if the following holds:

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

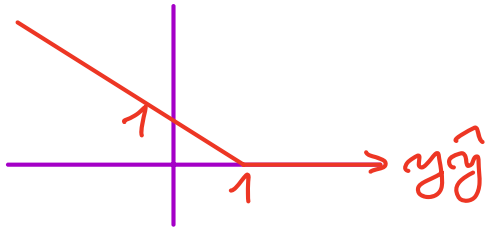
Thm: if ℓ is calibrated then if $\hat{y}(\cdot)$ is the measurable function minimizing R^{ℓ}, then it also minimizes $R^{0/1}$

Intuitively: A measurable function learnt to minimize a calibrated loss will also minimize the 0/1 loss

Calibration of scoring losses

Exercise:

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



- We will consider 3 cases in this exercise:
  - case 1: $\eta < \frac{1}{2}$
  - case 2: $\eta = 1/2$
  - case 3: $\eta > 1/2$
- Draw $C(\hat{y}, \eta)$ as a function of $\hat{y}$ in each of these three cases.
- In each case, show which value of $\hat{y}$ minimizes $C(\hat{y}, \eta)$
- Also show which predicted class corresponds to those $\hat{y}$
- If, instead of the hinge loss, we used the 0/1 loss, which class would be the optimal one in these 3 cases?
- Using the definition of calibration, show the hinge loss is calibrated

Calibration of scoring losses

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

$\Rightarrow$ all scoring losses we saw are well-calibrated