Fundamental of machine learning

Lecturer: Yann Chevaleyre
Scribe: ALJOHANI Renad

Lecture n°1
05/10/2023

1 Introduction

The ideal loss function, in the binary case, would be the 0/1 loss function, denoted as

$$l^{0/1}(\hat{y}, y) = 1(\hat{y} \neq y)$$

where it equals 1 if $\hat{y}$ is not equal to y (indicating a misclassification) and 0 otherwise. However, this loss function cannot be practically used primarily because it is non-convex.

To overcome the non-convexity issue, a common approach is to replace the 0/1 loss function with a convex surrogate loss function that maintains the essential characteristics of the learning problem while ensuring ease of optimization. This surrogate loss function must possess convexity to be suitable for optimization.

2 Linear classification

2.1 Linear classification with the 0/1 loss

let :

$$S = (x_1, y_1), (x_2, y_2), ..., (x_N, y_N), X \in \mathbb{R}^d, y = \hat{y} = \{0, 1\}$$

$$F = \{x \mapsto [\theta^T x + b \geq 0] : \theta \in \mathbb{R}^d, b \in \mathcal{R}\}$$

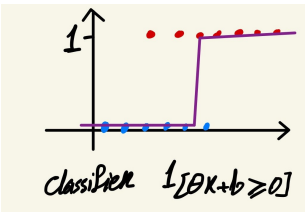
The learning problem is :

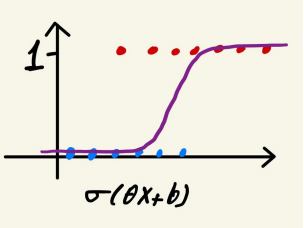
$$argmin_{f \in \mathcal{F}} \sum_{i=1}^{N} \ell^{0/1}\left(f(x_i), y_i\right) = \mathbf{1}_{[f(x_i) \neq y_i]} \Leftrightarrow argmin_{\theta, b} \sum_{h_i = y_i = 1} \mathbf{1}_{[\theta^T x_i + b < 0]} + \sum_{h_i = y_i = 1} \mathbf{1}_{[\theta^T x_i + b \geq 0]} \tag{1}$$

NB : easy problem if the classes are well seperated hard (Np-hard) otherwise.

2.2 Linear classification with the logistic regression framework

The main idea here is to replace the binary classifier with a continuous function, the parameters of which can be optimized through a convex optimization problem.





2.2.1 Linear classification with the logistic regression : Binary class setting

$y = \{0, 1\}$

Notation : (dependency on θ, b omitted for brevity).

— $\hat{y_i} = \theta^T x_i + b$ (the score of x_i)

— $\hat{\eta_i} = \sigma(\hat{y_i}) = \frac{1}{1+e^{-\hat{y_i}}}$ (the Class Probability Estimate CPE)

— $\hat{c_i} = \mathbf{1}_{[\hat{y_i} \geq 0]} = \mathbf{1}_{[\hat{\eta_i} \geq \frac{1}{2}]}$ (predicted class)

— $\hat{\eta_i}$ is interpreted as $P(y = 1|X = x_i, \theta, b)$

— $(1 - \hat{\eta_i})$ is interpreted as $P(y = 0|X = x_i, \theta, b)$

— pb : learn θ, b

Very naive idea : learn $\hat{\theta}, \hat{b} = argmin \sum_{i=1}^{N} \mathbf{1}_{[\hat{C_i} \neq y_i]}$ (Naive idea because identical to previous problem so hard)

Naive idea : learn $\hat{\theta}, \hat{b} = argmin \sum_{i=1}^{N} (\hat{\eta_i} - y_i)^2$ (Naive because the objective function is continous but not convex)

Exercise 1 : Show with a dataset of one example (0,0) that the objective function is not convex.

Solution :

— $argmin \sum_{i=1}^{N} (\hat{\eta_i} - y_i)^2$

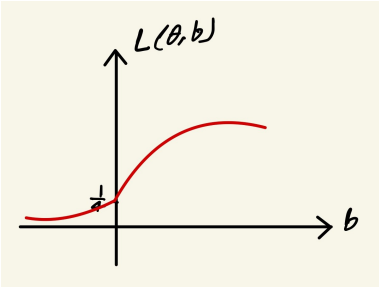
— recall : $L(\theta, b)$

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

$\hat{\eta_i} = \sigma(\hat{y_i}) = \frac{1}{1+e^{-\hat{y_i}}}$

— $x_1 = 0, y_1 = 0, \hat{y_1} = b, \hat{\eta_1} = \sigma(b) = \frac{1}{1+e^{-b}}$

$argmin_b \sigma(b)^2$



(so is non convex)

We will use the **probabilistic interpretation** of logistic regression to get a good loss. The likelihood of the model θ, b is :

$$\mathcal{L}(\theta, b) = \Pi_{i=1}^{N} p(y = y_i|X = x_i, \theta, b) = \pi_{i:y_i=1} \hat{\eta_i} \cdot \pi_{i:y_i=0} (1 - \hat{\eta_i}) \tag{2}$$

Making the data the most probable with regard to θ, b amounts to minimizing $\alpha(\theta, b)$ or

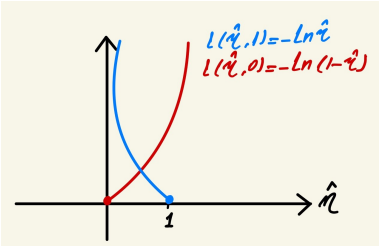
to minimize the negative log likelihood $\mathcal{NL}\alpha(\theta, b)$.

$$\begin{aligned}\mathcal{NL}\alpha(\theta, b) &= -\log \mathcal{L}(\theta, b) \\ &= -\sum_{i:y_i=1} \ln \hat{\eta}_i - \sum_{i:y_i=0} \ln(1-\hat{\eta}_i) \\ &= \sum_{i=0}^{N} -y_i \ln \hat{\eta}_i - (1-y_i)\ln(1-\hat{\eta}_i) \\ &= l^{CE}(\hat{\eta}_i, y_i) \quad \textbf{(cross-entropy loss)}\end{aligned} \tag{3}$$

The logistic regression problem is :

$$\hat{\theta}, \hat{b} = argmin_{\theta,b} \sum_{i=1}^{N} l^{CE}(\hat{\eta}_i, y_i) \tag{4}$$

Exercise 2 :
— Draw $l^{CE}(\hat{\eta}, 1)$ and $l^{CE}(\hat{\eta}, 0)$
— Based on the drawing, show $argmin_{\hat{\eta}} l^{CE}(\hat{\eta}, 0)$ and $argmin_{\hat{\eta}} l^{CE}(\hat{\eta}, 1)$.
Solution :
$l^{CE}(\hat{\eta}, 0) = -\ln(1-\hat{\eta})$
$l^{CE}(\hat{\eta}, 1) = -\ln \hat{\eta}$



$argmin_{\hat{\eta}} l^{CE}(\hat{\eta}, 0) = 0$
$argmin_{\hat{\eta}} l^{CE}(\hat{\eta}, 1) = 1$
note : $\hat{\eta}$ never negative because it is a probability.

2.2.2 Linear classification with the logistic regression : the multiclass setting

$y = \{1..k\}$
Notation :
— foreach $k \in y$, we have $\theta_k \in \mathbb{R}^d, b_k \in \mathbb{R}^d$
— $(\hat{y}_{i,1}, ..., \hat{y}_{i,K}) = (\theta_1^T x_i + b_1, ..., \theta_K^T x_i + b_K)$ (the score of x_i)
— replace sigmoid by softmax : $softmax \begin{pmatrix} t_1 \\ . \\ . \\ t_k \end{pmatrix} = \frac{1}{\sum_{k=1}^{K} e^{t_k}} \begin{pmatrix} e^{t_1} \\ . \\ . \\ e^{t_K} \end{pmatrix}$
— $\hat{\eta}_i = (\hat{\eta}_{i,1}, ..., \hat{\eta}_{i,K}) = softmax(\hat{y}_{i,1}, ..., \hat{y}_{i,K})$ (the Class Probability Estimate CPE)
— $\hat{c}_i = argmax_{k \in \{1..k\}} \hat{y}_{ik}$ (predicted class)

— multiclass cross-entropy loss $l^{CE}(\hat{\eta}_i, y_i) = \sum_{k=1}^{K} -\mathbf{1}_{[y_i=k]} \ln \hat{\eta}_{i,k}$

Example for multiclass setting :
we have three classes : $y = \{\text{dog, cat, mouse}\}$ and $X = (2, 3)$

Parameters		
$\theta_{\text{dog}} = (1, 1)^T$	$\theta_{\text{cat}} = (-1, 0)^T$	$\theta_{\text{mouse}} = (1, -1)^T$
$b_{\text{dog}} = 0$	$b_{\text{cat}} = 0$	$b_{\text{mouse}} = 0$
Scores		
$\hat{y}_{\text{dog}} = 5$	$\hat{y}_{\text{cat}} = -2$	$\hat{y}_{\text{mouse}} = -1$
CPEs		
$\exp(\hat{y}_{dog}) = 148$	$\exp(\hat{y}_{cat}) = 0.1$	$\exp(\hat{y}_{mouse}) = 0.4$
$\hat{\eta}_{\text{dog}} = 99.6\%$	$\hat{\eta}_{\text{cat}} = 0.1\%$	$\hat{\eta}_{\text{mouse}} = 0.2\%$
Predicted Class : $\hat{c} = \text{dog}$		

2.3 The two views of logistic regression : (in the binary class setting)

View 1 : Working with cross-entropy loss $l^{CE}(\hat{\eta}, y)$ with $y = \{0, 1\}$ which is a CPE-loss.

$$\hat{\theta}, \hat{b} = argmin_{\theta,b} \sum_{i=1}^{N} l^{CE}(\hat{\eta}_i, y_i) \tag{5}$$

view 2 : Work with the logistic loss $l^{logistic}(\hat{y}, y)$ with $y = \{-1, 1\}$ whic is a scoring loss.

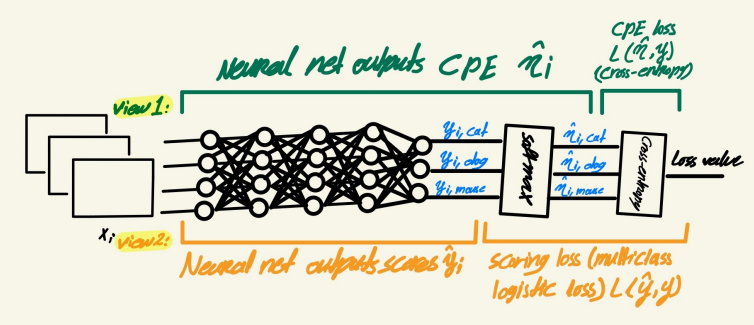
$$\hat{\theta}, \hat{b} = argmin_{\theta,b} \sum_{i=1}^{N} l^{logistic}(\hat{y}_i, y_i) \tag{6}$$

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

Both views are identical, and are convex in θ, b
Exercise 3 : Developp view 1 to show it is identical to view2.
Solution :

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

In the analysis of a Neural Network for binary classification, there exist two approaches for decomposition : one using cross-entropy loss and the other using scoring loss. It is noteworthy that both the cross-entropy loss function and scoring functions can be effectively employed in binary classification tasks. As demonstrated in the following example, we trained a neural network on 64x64 images with the class labels $y = \{cat, dog, mouse\}$ and the feature space $X = \mathbb{R}^{64x64}$.



3 CPE-Losses (Class probability Estimate Losses

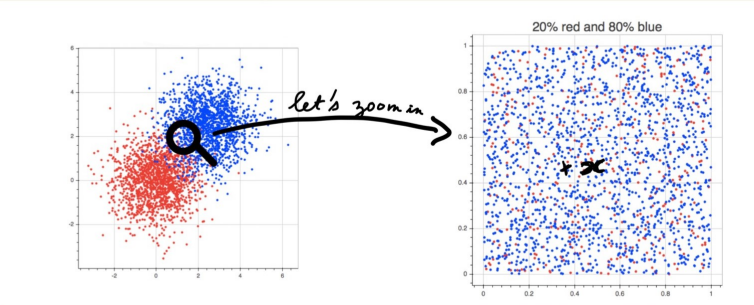
In this section, we restrict to $y = 0, 1$. Recall that CPE loss is of the form $l(\hat{\eta}, y)$ (e.g. cross entropy).

— What are good CPE-losses ?

— **Example 1 :** Assuming in my dataset, I have a bunch of images indistinguishable by NN. Among these images, I have 80% mouse and 20% cats. What should the NN output on these images ?

It should output : $\hat{\eta}_{i,mouse} = 0,8 \quad \hat{\eta}_{i,cat} = 0,2$

— **Example 2 :** $y = \{0, 1\}$



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 included by points in this region will be $\eta l(\hat{\eta}, 1) + (1 - \eta)l(\hat{\eta}, 0)$

3.1 Risk of CPE-loss

Notation :

— $\hat{\eta} \in [0, 1]$ is a probability.

— $\hat{\eta}(.) \in \mathbb{R}^d \mapsto [0, 1]$ is a CPE prediction function (e.g. a neural network with softmax.

let us first adapt our definition of the risk to new loss.

Definition 1. *The risk of a prediction function $\hat{\eta}(.)$ with CPE-loss l is :*

$$\mathcal{R}^l(\hat{\eta}(.)) = \mathbb{E}_{x,y\sim p}[l(\hat{\eta}(x), Y)] \tag{7}$$

To compute the intuition we need the conditional risk.

3.1.1 Conditional risk of CPE-loss

Definition 2. *For $\eta \in [0,1]$ and $\hat{\eta} \in [0,1]$, the conditional risk is :*

$$\mathcal{C}(\hat{\eta}, \eta) = \mathbb{E}_{y \sim Ber(\eta)}[l(\hat{\eta}, Y)] = \eta(x).l(\hat{\eta}(x), 1) + (1 - \eta(x)).l(\hat{\eta}(x), 0) \tag{8}$$

$\mathcal{R}^l(\hat{\eta}(.)) = \mathbb{E}_{x,y \sim p}[l(\hat{\eta}(x), Y)]$

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

$\mathbb{E}_x[\eta(x).l(\hat{\eta}(x), 1) + (1 - \eta(x)).l(\hat{\eta}(x), 0)]$

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

We can discern the relationship between conditional risk and risk as described by the equations :

$$\mathcal{R}^l(\hat{\eta}(x)) = \mathbb{E}_x y(l\hat{\eta}(x), y) = \mathbb{E}_x \mathbb{E}_{Y|x}[l(\hat{\eta}(x), y] = \mathbb{E}_x \mathcal{C}(\hat{\eta}(x), \eta(x)) \tag{9}$$

where $\eta(x) = P(Y = 1|X = x)$

3.2 Proper CPE losses

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

Definition 3. $\forall \eta \in [0,1] \quad \eta \in argmin_{\hat{\eta}} \mathcal{C}(\hat{\eta}, \eta)$

Exercise 4 :
Exercise 4.1 : Show the cross entropy is proper.
Solution :
let us we show that $\eta = argmin_{\hat{\eta}} \mathcal{C}(\hat{\eta}, \eta)$

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

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

$\eta(1 - \hat{\eta}) = \hat{\eta}(1 - \eta) \Rightarrow \eta = \hat{\eta}$ (η is equal to the $\hat{\eta}$ which minimize $\mathcal{C}(\hat{\eta}, \eta)$)

l^{CE} is (strictly) proper

Exercise 4.2 : Write the Bayes risk ($= min_f \mathbb{R}^l(f)$) for the case l is proper. Apply the cross entropy.
Solution :
$inf_f \mathcal{R}^l(f) = \mathbb{E}_x inf_{\hat{\eta}} \mathcal{C}(\hat{\eta}, \eta(x)) = \mathbb{E}_x \mathcal{C}(\eta(x), \eta(x))$
with cross-entropy, $\mathbb{E}_x \mathcal{C}(\eta(x), \eta(x)) = \mathbb{E}_x H(\eta(x))$ where $\mathcal{H}(.)$ is entropy.

3.3 More on CPE-losses and the cross-entropy

— $l(\hat{\eta}, y)$ where $\hat{\eta}$ is a parameter of a Bernoulli distribution, y in $\{0, 1\}$ could also be interpreted as a parameter of Bernouilli. For example : $l(10\%, 0)$ can be interpreted as a dissimilarity between two distribution $Ber(10\%)$ and $Ber(0\%)$.

— There is a standard way to compare discrete distributions : Kullback-Leibler (KL) divergence for two distributions p and q on $\mathcal{Y}$,

$$KL(p||q) = \sum_{y \in \mathcal{Y}} p(y) \ln \frac{p(y)}{q(y)} \tag{10}$$

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

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

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

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

(Assumption $0\ln 0 = 0$)

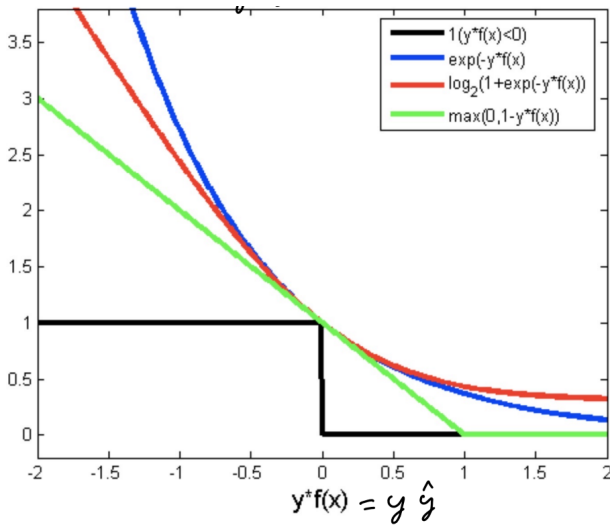
$l^{CE}(\hat{\eta}, y) = KL(Ber(y), Ber(\hat{\eta}))$

4 Scoring losses

let $l(\hat{y}, y)$ be a loss function where $\hat{y}_i = \theta^T x + b$ suppose there some scoring losses such that $l(\hat{y}, y) = \phi(\hat{y}, y)$. This loss is called ϕ-losses or (margin losses). For example :

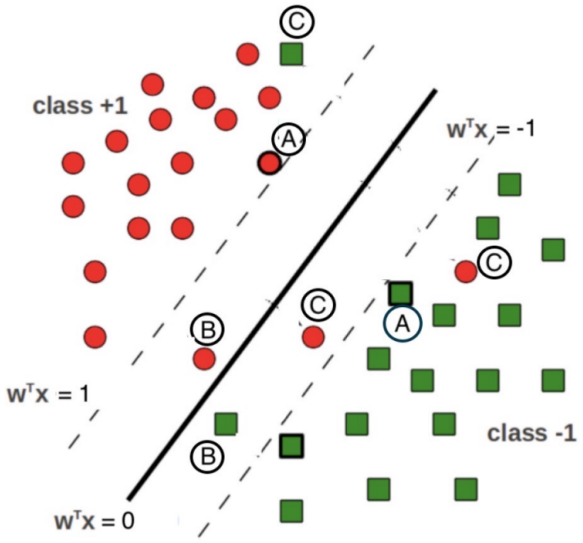
— $0/1$ loss : $l^{0/1}(\hat{y}, y) = 1[y\hat{y} < 0]$
— hinge loss : $l^{hinge}(\hat{y}, y) = max(1 - y\hat{y}, 0)$
— logistic loss : $l^{logistic}(\hat{y}, y) = \ln(1 + e^{-y\hat{y}})$

Several of these loss functions are upper bounds of the $0/1$ loss as shown on the following graph.



In the figure above, a comparison is made with the $0/1$ loss. These margin losses take into account the distance between the predicted value $\hat{y}$ and the actual class y.

note :$y\hat{y}$ must be positive, if negative then not well classified.



(A) Points properly classified
(B) Points at the margin yet properly classified
(C) Points improperly classified

4.0.1 The risk of a scoring loss

Notations :
— $\eta(x) = P(Y = 1|X = x)$
— $\hat{y} \in \mathcal{R} \cup \{-\infty, \infty\}$ is a score
— $\hat{y}(.) \in \mathcal{X} \mapsto \mathcal{R} \cup \{-\infty, \infty\}$ is a scoring function (e.g. neural net).

Definition 4. *For any $\hat{y} \in \mathcal{R}$, the conditional risk for scoring loss l is*

$$\begin{aligned} \mathcal{C}(\hat{y}, \eta) &= \mathbb{E}_{y \sim \begin{cases} 1 \ w.p. \ \eta \\ -1 \ w.p. \ 1-\eta \end{cases}} [l(\hat{y}, y)] \\ &= \eta l(\hat{y}, 1) + (1-\eta) l(\hat{y}, -1) \end{aligned} \tag{11}$$

Definition 5. *The risk of scoring function $\hat{y}(.) : \mathcal{R}^d \to \mathcal{R}$ for loss l is*

$$\mathcal{R}^l(\hat{y}(.)) = \mathbb{E}_{x,y \sim p}[l(\hat{y}(x), y)] = \mathbb{E}_x C(\hat{y}(x), \eta(x)) \tag{12}$$

4.0.2 Calibration of scoring losses

What are "good" scoring losses ?

Definition 6. *A loss l is calibrated if the following holds :*
— *if $\eta \in [0, \frac{1}{2}[$ then $inf_{\hat{y}<0} \mathcal{C}(\hat{y}, \eta) < inf_{\hat{y} \geq 0} \mathcal{C}(\hat{y}, \eta)$*
— *if $\eta \in]\frac{1}{2}, 1]$ then $inf_{\hat{y}>0} \mathcal{C}(\hat{y}, \eta) < inf_{\hat{y} \leq 0} \mathcal{C}(\hat{y}, \eta)$*

Theorem 1. *If a loss function l is calibrated, and if $\hat{y}(.)$ is the measurable function that minimizes $\mathcal{R}^l$, then it also minimizes $\mathcal{R}^{l^{0,1}}$*

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

Theorem 2. *Any convex $\phi-$loss with $\phi < 0$ is well calibrated*

5 Conclusion

In this lecture, we've explored the world of loss functions and their significance in machine learning and classification tasks. We've delved into the two major categories of loss functions : Class Probability Estimate (CPE) losses and scoring losses.

For CPE losses, we've discussed proper loss functions that yield well-calibrated results, and we've provided insights into their relevance in various classification scenarios. We've particularly emphasized the proper use of cross-entropy loss and its applications in training neural networks for binary classification.

Scoring losses, on the other hand, offer an alternative perspective for evaluating classification models. We've presented different scoring losses, such as the $0/1$ loss, hinge loss, and logistic loss, highlighting their role in taking into account the margin or distance between predicted values and actual class labels. Importantly, it's worth noting that all scoring losses we've examined in this lecture are well-calibrated.