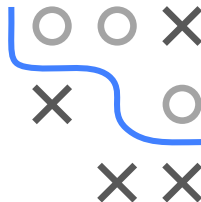


COST-SENSITIVE LEARNING: IN A NUTSHELL

- Cost-sensitive learning:
 - Classical learning: data sets are balanced, and all errors have equal costs
 - We now assume given, unequal cost
 - And try to minimize them in expectation
- Applications:
 - Medicine — Misdiagnosing as healthy vs. having a disease
 - (Extreme) Weather prediction — Incorrectly predicting that no hurricane occurs
 - Credit granting — Lending to a risky client vs. not lending to a trustworthy client.



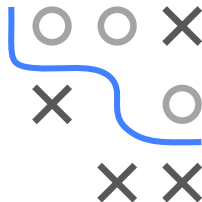
		Truth	
		Default	Pays Back
Pred.	Default	0	10
	Pays Back	1000	0

- In these examples, **the costs of a false negative is much higher than the costs of a false positive.**
- In some applications, the costs are **unknown** \rightsquigarrow need to be specified by experts, or be learnt.

COST MATRIX

- Input: cost matrix \mathbf{C}

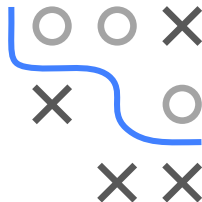
		True Class y			
		1	2	...	g
Classification	1	$C(1, 1)$	$C(1, 2)$...	$C(1, g)$
	\hat{y}	2	$C(2, 1)$	$C(2, 2)$...
	\vdots	\vdots	\vdots	...	\vdots
	g	$C(g, 1)$	$C(g, 2)$...	$C(g, g)$



- $C(j, k)$ is the cost of classifying class k as j ,
- 0-1-loss would simply be: $C(j, k) = \mathbb{1}_{[j \neq k]}$
- \mathbf{C} designed by experts with domain knowledge
 - 1 Too low costs: not enough change in model, still costly errors
 - 2 Too high costs: might never predict costly classes

COST MATRIX FOR IMBALANCED LEARNING

- Common heuristic for imbalanced data sets:
 - $C(j, k) = \frac{n_j}{n_k}$ with $n_k \ll n_j$,
misclassifying a minority class k as a majority class j
 - $C(j, k) = 1$ with $n_j \ll n_k$,
misclassifying a majority class k as a minority class j
 - 0 for a correct classification



- Imbalanced binary classification:

	True class	
	$y = 1$	$y = -1$
Pred. $\hat{y} = 1$	0	1
class $\hat{y} = -1$	$\frac{n_-}{n_+}$	0

- So: much higher costs for FNs

MINIMUM EXPECTED COST PRINCIPLE

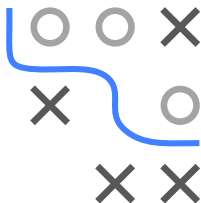
- Suppose we have:
 - a cost matrix \mathbf{C}
 - knowledge of the true posterior $p(\cdot | \mathbf{x})$
- Predict class j with smallest expected costs when marginalizing over true classes:

$$\mathbb{E}_{K \sim p(\cdot | \mathbf{x})}(C(j, K)) = \sum_{k=1}^g p(k | \mathbf{x}) C(j, k)$$

- If we trust we trust a probabilistic classifier, we can convert its scores to labels:

$$h(\mathbf{x}) := \arg \min_{j=1, \dots, g} \sum_{k=1}^g \pi_k(\mathbf{x}) C(j, k).$$

- Can be better to take a less probable class ([Elkan et. al. 2001](#))



OPTIMAL THRESHOLD FOR BINARY CASE

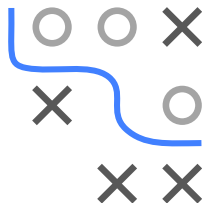
- Optimal decisions do not change if
 - \mathbf{C} is multiplied by positive constant
 - \mathbf{C} is added with constant shift
- Scale and shift \mathbf{C} to get simpler \mathbf{C}' :

	True class	
	$y = 1$	$y = -1$
Pred. $\hat{y} = 1$	$C'(1, 1)$	1
class $\hat{y} = -1$	$C'(-1, 1)$	0

where

- $C'(-1, 1) = \frac{C(-1,1)-C(-1,-1)}{C(1,-1)-C(-1,-1)}$
- $C'(1, 1) = \frac{C(1,1)-C(-1,-1)}{C(1,-1)-C(-1,-1)}$
- We predict \mathbf{x} as class 1 if

$$\mathbb{E}_{K \sim p(\cdot | \mathbf{x})}(C'(1, K)) \leq \mathbb{E}_{K \sim p(\cdot | \mathbf{x})}(C'(-1, K))$$



OPTIMAL THRESHOLD FOR BINARY CASE / 2

- Let's unroll the expected value and use C' :

$$p(-1 | \mathbf{x})C'(1, -1) + p(1 | \mathbf{x})C'(1, 1) \leq p(-1 | \mathbf{x})C'(-1, -1) + p(1 | \mathbf{x})C'(-1, 1)$$

$$\Rightarrow [1 - p(1 | \mathbf{x})] \cdot 1 + p(1 | \mathbf{x})C'(1, 1) \leq p(1 | \mathbf{x})C'(-1, 1)$$

$$\Rightarrow p(1 | \mathbf{x}) \geq \frac{1}{C'(-1, 1) - C'(1, 1) + 1}$$

$$\Rightarrow p(1 | \mathbf{x}) \geq \frac{C(1, -1) - C(-1, -1)}{C(-1, 1) - C(1, 1) + C(1, -1) - C(-1, -1)} = c^*$$

- If even $C(1, 1) = C(-1, -1) = 0$, we get:

$$p(1 | \mathbf{x}) \geq \frac{C(1, -1)}{C(-1, 1) + C(1, -1)} = c^*$$

- Optimal threshold c^* for probabilistic classifier

$$h(\mathbf{x}) := 2 \cdot \mathbb{1}_{[\pi(\mathbf{x}) \geq c^*]} - 1$$

