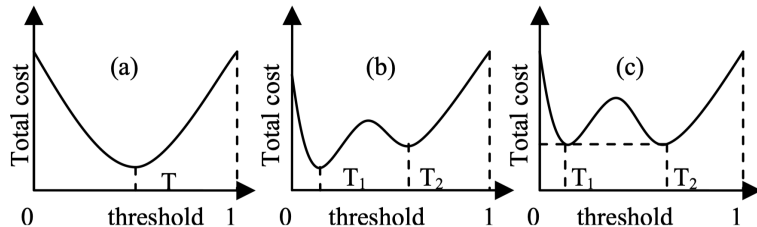
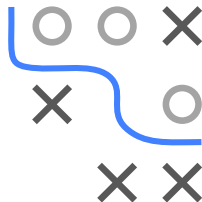


EMPIRICAL THRESHOLDING: BINARY CASE

- Theoretical threshold from MECP not always best, due to e.g. wrong model class, finite data, etc.
- Simply measure costs on data with different thresholds
- Then pick best threshold (Fig.1 in [Sheng et al. 2006](#)):



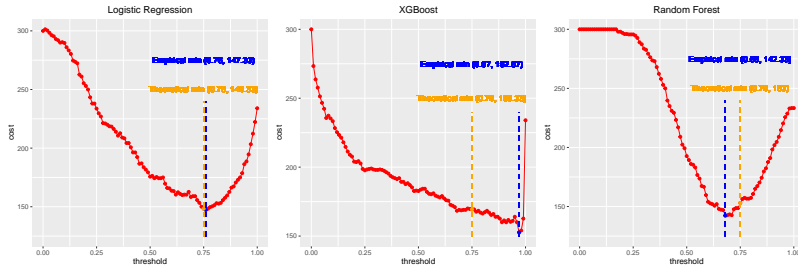
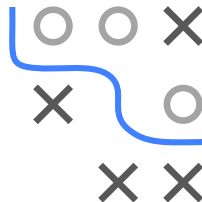
- What if two equal local minima? We prefer the one with wider span
- Do this on validation data / over cross-val to avoid overfitting!

EMPIRICAL THRESHOLDING: BINARY CASE

- Example: German Credit task

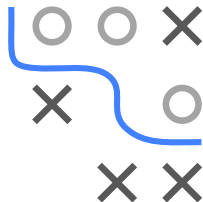
	True class	
	$y = \text{good}$	$y = \text{bad}$
Pred. $\hat{y} = \text{good}$	0	3
class $\hat{y} = \text{bad}$	1	0

- Theoretical: $C(\text{good}, \text{bad}) / (C(\text{bad}, \text{good}) + C(\text{good}, \text{bad})) = 3/4 = c^*$
- Empirical version with 3-CV: For XGBoost, empirical minimum deviates substantially from theoretical version



EMPIRICAL THRESHOLDING: MULTICLASS

- In the standard setting, we predict class $h(\mathbf{x}) = \arg \max_k \pi_k(\mathbf{x})$.
- Let's use g thresholds c_k now
- Re-scale scores $\mathbf{s} = (\frac{\pi(\mathbf{x})_1}{c_1}, \dots, \frac{\pi(\mathbf{x})_g}{c_g})^\top$,
- Predict class $\arg \max_k \pi_k(\mathbf{x})$.
- Compute empirical costs over cross-validation
- Optimize over g (actually: $g - 1$) dimensional threshold vector $(c_1, \dots, c_g)^\top$ to produce minimal costs



METACOST: ALGORITHM

Input: $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n$ training data, B number of bagging iterations, $\pi(\mathbf{x})$ probabilistic classifier, \mathbf{C} cost matrix, empty dataset $\tilde{\mathcal{D}} = \emptyset$

Bagging: Classifier is trained on different bootstrap samples.

for $b = 1, \dots, B$ **do**

$\mathcal{D}_b \leftarrow$ Bootstrap version of \mathcal{D}

$\pi_b \leftarrow$ train classifier on \mathcal{D}_b

end for

Relabeling: Find classifiers for which $\mathbf{x}^{(i)}$ is OOB and compute π_b by averaging over predictions. Determine new label $\tilde{y}^{(i)}$ w.r.t. to the cost minimal class.

for $i = 1, \dots, n$ **do**

$\tilde{M} \leftarrow \bigcup_{m: \mathbf{x}^{(i)} \notin \mathcal{D}_m} \{m\}$

end for

for $j = 1, \dots, g$ **do**

$\pi_j(\mathbf{x}^{(i)}) \leftarrow \frac{1}{|\tilde{M}|} \sum_{m \in \tilde{M}} \pi_j(\mathbf{x}^{(i)} \mid f_m)$ for each i

end for

for $i = 1, \dots, n$ **do**

$\tilde{y}^{(i)} \leftarrow \arg \min_k \sum_{j=1}^g \pi_j(\mathbf{x}^{(i)}) C(k, j)$

$\tilde{\mathcal{D}} \leftarrow \tilde{\mathcal{D}} \cup \{(\mathbf{x}^{(i)}, \tilde{y}^{(i)})\}$

end for

Cost Sensitivity: Train on relabeled data.

$f_{\text{meta}} \leftarrow$ train f on $\tilde{\mathcal{D}}$

