Performance metrics for supervised learning

Algorithms & Data Challenges Berlin
Talks n' Beer IV Meetup

Alexander Weiß

October 9, 2013

Shock your CEO, read a scientific paper!

A Unified View of Performance Metrics: Translating Threshold Choice into Expected Classification Loss. J. Hernández-Orallo, P. Flach, and C. Ferri. J. Mach. Learn. Res. 13(1). January 2012.

The question I - binary classifier

binary classifier: $X \to \{0, 1\}$

0 - negative
1 - positive

naive Bayes, perceptron, SVM, ...

The question I - binary classifier vs. probabilistic model

binary classifier: $X \to \{0, 1\}$

0 - negative
1 - positive

naive Bayes, perceptron, SVM, ...

---

probabilistic model: $m : X \to [0, 1]$

$m(x) \approx \mathbb{P}(positive|x)$

logistic regression, Probabilities for SVMs (J. Platt, 1999), ...

The question II - operating conditions

From probabilistic model to classifier:

Choose $T \in [0, 1]$. If $m(x) \geq T$, predict 1, else 0

The question II - operating conditions

From probabilistic model to classifier:

$$\text{Choose } T \in [0,1]. \text{ If } m(x) \geq T, \text{ predict 1, else 0}$$

$T : \Theta \to [0,1]$ is a function of the operating conditions. For $\theta \in \Theta$:

$$\theta = (b, c, \pi_0).$$

- c_k is cost of misclassification of k-sample, $k \in [0,1]$
- $b := c_0 + c_1$ is overall cost
- $c := c_0/b$ is cost fraction of negative sample
- π_k is fraction of k-samples in X, $k \in [0,1]$; $\pi_0 + \pi_1 = 1$

What's the performance of a model?

A lot of metrics available

- accuracy (for a threshold t):
  $Acc(t) := \pi_0 \; F_0(t) + \pi_1 \; (1 - F_1(t))$

What's the performance of a model?

A lot of metrics available

- accuracy (for a threshold t):
  $Acc(t) := \pi_0 \; F_0(t) + \pi_1 \; (1 - F_1(t))$
- Brier score:
  $\int_0^1 \left(f_0(s) s^2 + f_1(s)(1-s)^2 \right) ds$

What's the performance of a model?

A lot of metrics available

- accuracy (for a threshold t):
  $Acc(t) := \pi_0 \; F_0(t) + \pi_1 \; (1 - F_1(t))$
- Brier score:
  $\int_0^1 \left(f_0(s) s^2 + f_1(s)(1-s)^2 \right) ds$
- Area under ROC curve:
  ROC: $[0,1] \to [0,1]^2$
  $\text{ROC}(t) := (\int_t^1 f_0(s) ds, \int_t^1 f_1(s) ds)$

What's the performance of a model?

A lot of metrics available

- accuracy (for a threshold t):
  $Acc(t) := \pi_0 \; F_0(t) + \pi_1 \; (1 - F_1(t))$
- Brier score:
  $\int_0^1 \left(f_0(s)s^2 + f_1(s)(1-s)^2 \right) ds$
- Area under ROC curve:
  ROC: $[0,1] \to [0,1]^2$
  $\text{ROC}(t) := (\int_t^1 f_0(s)ds, \int_t^1 f_1(s)ds)$
  ROC(t) = (false pos. rate(t), true pos. rate(t))

What's the performance of a model?

A lot of metrics available

- accuracy (for a threshold t):
  $Acc(t) := \pi_0 \; F_0(t) + \pi_1 \; (1 - F_1(t))$
- Brier score:
  $\int_0^1 \left(f_0(s) s^2 + f_1(s)(1-s)^2 \right) ds$
- Area under ROC curve:
  ROC: $[0,1] \to [0,1]^2$
  $\text{ROC}(t) := (\int_t^1 f_0(s)ds, \int_t^1 f_1(s)ds)$
  ROC(t) = (false pos. rate(t), true pos. rate(t))

They're not consistent!

Define your operating conditions!

$T : \Theta \to [0, 1]$ is a function of the operating conditions. For $\theta \in \Theta$:

$$\theta = (b, c, \pi_0).$$

- c_k is cost of misclassification of k-sample, $k \in [0, 1]$
- $b := c_0 + c_1$ is overall cost
- $c := c_0/b$ is cost fraction of negative sample
- π_k is fraction of k-samples in X, $k \in [0, 1]$; $\pi_0 + \pi_1 = 1$

Simplification for the talk

- overall cost b will be 1 in expectation: $\mathbb{E}b = 1$
- cost fraction c will be random variable with distribution w
- fraction of k-samples π_k will be constant

Simplification for the talk

- overall cost b will be 1 in expectation: $\mathbb{E}b = 1$
- cost fraction c will be random variable with distribution w
- fraction of k-samples π_k will be constant

Expected loss is given by

$$L_c(t) = \int_0^1 \left(c\pi_0(1 - F_0(t)) + (1 - c)\pi_1 F_1(t)\right) w(c)dc$$

Answer I: A score-fixed threshold

Let the threshold

$$T^{sf}(\theta) = T^{sf}(c) := \tilde{t}$$

be fixed.

Answer I: A score-fixed threshold

Let the threshold

$$T^{sf}(\theta) = T^{sf}(c) := \tilde{t}$$

be fixed.

Theorem (for score-fixed thresholds)

If a classifier sets the decision threshold at a fixed value $\tilde{t}$ irrespective of the operating condition or the model, then expected loss for any cost distribution w is given by:

$$L_c^{sf}(\tilde{t}) = 2\mathbb{E}_w(c)(1 - Acc(\tilde{t})) + 4\pi_1 F_1(\tilde{t}) \left(\frac{1}{2} - \mathbb{E}_w(c) \right)$$

Answer II: A score-driven threshold

Let the threshold be

$$T^{sd}(\theta) = T^{sd}(c) := c.$$

Answer II: A score-driven threshold

Let the threshold be

$$T^{sd}(\theta) = T^{sd}(c) := c.$$

Theorem (for score-driven thresholds, Hernández-Orallo et al. (2011))

Assuming probabilistic score and the score-driven threshold choice method, exptected loss under a uniform distribution of cost proportions is equal to the model's Brier score.

Answer III: A rate-driven threshold

Let $R(t)$ be the fraction of samples predicted as negative:

$$R(t) := \pi_0 F_0(t) + \pi_1 F_1(t)$$

Answer III: A rate-driven threshold

Let $R(t)$ be the fraction of samples predicted as negative:

$$R(t) := \pi_0 F_0(t) + \pi_1 F_1(t)$$

Let the threshold

$$T^{rd}(\theta) = T^{rd}(c) := R^{-1}(c)$$

Answer III: A rate-driven threshold

Let $R(t)$ be the fraction of samples predicted as negative:

$$R(t) := \pi_0 F_0(t) + \pi_1 F_1(t)$$

Let the threshold

$$T^{rd}(\theta) = T^{rd}(c) := R^{-1}(c)$$

Theorem (for rate-driven thresholds)

Expected loss for uniform cost proportions using the rate-driven threshod choice method is linearly related to AUC as follows:

$$L^{rd}_{U(c)} = \pi_0 \pi_1 (1 - 2AUC) + 1/3$$

Take home messages

1. There is no model naturally good or bad; its performance depends on the operating conditions and the way the threshold is chosen.

Take home messages

1. There is no model naturally good or bad; its performance depends on the operating conditions and the way the threshold is chosen.
2. If you know the threshold method, you know the right performance metric to choose the best model.

Take home messages

1. There is no model naturally good or bad; its performance depends on the operating conditions and the way the threshold is chosen.
2. If you know the threshold method, you know the right performance metric to choose the best model.
3. Read the paper. It's much more detailed and much more extensive.

Take home messages

1. There is no model naturally good or bad; its performance depends on the operating conditions and the way the threshold is chosen.
2. If you know the threshold method, you know the right performance metric to choose the best model.
3. Read the paper. It's much more detailed and much more extensive.
4. If you have questions, ask now. It's too late when you're at home.