Proper Scoring Rules

David S. Rosenberg

NYU: CDS

November 17, 2021

Contents

Introduction and Motivation

Binary prediction: hard and soft

- Methods like SVMs don't naturally give probability predictions
  - How can we get probability predictions from SVM predictions?
- Most methods these days do produce probabilities naturally
  - logistic regression
  - gradient boosting with logistic loss
  - neural networks with logistic loss (called various other things in NN context)
- But how reliable are these probabilities as probabilities?
- We'll investigate this in the next module on calibrated probability predictions.
- In this short module, we review basic measures of prediction quality for probabilities.

- Many common machine learning models produce probability distributions over outputs for any given input. Common examples include logistic regression, multinomial logistic regression, Poisson regression, and any neural network model that produces the output of a softmax function. By far the most common way to train these models is to use maximum likelihood (often referred to as cross-entropy loss in the neural network literature). But what is the best way to evaluate these models?
- Of course, the best way to evaluate depends on what we're doing with the model downstream. If the probability prediction doesn't matter per se, e.g. we just want to rank people by how likely they are to be X, then standard metrics such as AUC-ROC and Precision@K may make sense. But if the probability distribution itself is the object of interest, for example if we're going to be sampling from these distributions in some downstream computations, then we need a way to evaluate the quality of these predictions. A "scoring rule" is a way to measure the "accuracy" (in some sense) of a predicted probability distribution.

Scoring rules: loss functions for probability forecasts

Predicting rain

- We want to predict the probability of rain tomorrow.
  - One forecaster says 70% chance.
  - Another forecaster says 60% chance.
  - It ends up raining – whose forecast is better?

If forced, we'd probably say the 70% forecast was better, because it gave higher probability to the actual outcome. But it seems too much to rank forecasters based on a single outcome. IF somehow we knew the true probability of rain (pretending / assuming nature decided based on a draw from a Bernoulli distribution), then we would prefer the forecast that's closest to the true probability, in some sense (e.g. square distance between the probabilities, or Kullback-Leibler divergence between the corresponding probability distributions).

Scoring rules – high level, general case

- Let P be predicted distribution on some set of outcomes $\mathcal{Y}$.
  - e.g. $\mathcal{Y} = \{\text{RAIN}, \text{NO RAIN}\}$
  - Let $y \in \mathcal{Y}$ be the actual outcome.
  - A scoring rule evaluates how good P in light of outcome y.
  - We'll write $S(P, y)$ for the evaluation of scoring rule S on P and y.
  - Conventions vary, but we'll assume smaller score is better.
  - This makes scoring rules special cases of loss functions.

Expected score

- Suppose we predict distribution P and the true distribution is Q.
- Then the expected score is written as $S(P, Q)$, where

$$S(P, Q) := \sum_{y \in \mathcal{Y}} Q(Y = y) S(P, y),$$

  where, for simplicity, we've assumed $\mathcal{Y}$ is a discrete set of outcomes.
- We would like to make a prediction P that optimizes $S(P, Q)$.

What's a good scoring rule?

- We want to optimize our expected score.
- Suppose somehow we actually knew the true distribution Q.
- So we should definitely just predict $Q\ldots$ right?
- Well... we should predict P where

$$P = \arg\min_{P} S(P, Q).$$

- Will Q always optimize the score when Q is the true distribution?
- This seems like something we want for a scoring rule...

Proper scoring rules

Definition (Proper scoring rule)

We say a scoring rule is proper if

$$S(Q, Q) \leqslant S(P, Q)$$

for all P and Q.

With a proper scoring rule, predict what you believe

Suppose our best guess for Q is $\hat{Q}$. If S is a proper scoring rule, then

$$S(\hat{Q}, \hat{Q}) \leqslant S(P, \hat{Q})$$

for all P. So our personal expectation of score is minimized when we predict $\hat{Q}$.

Proper scoring rule example: negative log-likelihood

- Suppose P can be represented by a PDF or PMF p, then we can define a scoring rule as

$$S(p, y) = -\log p(y).$$

- Is this a proper scoring rule?Yes!
- The expected score for predicting p for Y when true distribution is q is

$$\begin{aligned} S(p, q) = \mathbb{E}_{Y \sim q}(-\log p(Y)) &= \mathbb{E}_{Y \sim q}\left(\log\left[\frac{q(Y)}{p(Y)}\frac{1}{q(Y)}\right]\right) \\ &= \mathbb{E}_{Y \sim q}\left(\log\left[\frac{q(Y)}{p(Y)}\right]\right) + \mathbb{E}_{Y \sim q}\left(\log\left[\frac{1}{q(Y)}\right]\right) \\ &= \mathrm{KL}(q \| p) + H(q) \end{aligned}$$

- By Gibbs inequality, $\mathrm{KL}(q \| p)$ is minimized when $p = q$.

Proper scoring rule example: Brier score (binary case)

- Suppose we are forecasting the probability of an event.
- The Brier score (binary version) for predicted probability p is defined by

$$S(p, y) = (p - y)^2,$$

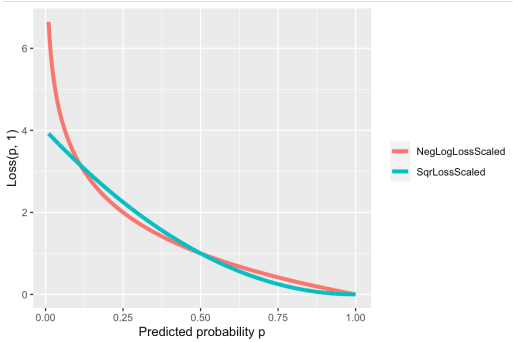
  where $y \in \{0, 1\}$.
- If $\mathbb{P}(Y = 1) = q$, then the expected score for predicting p is

$$\begin{aligned} S(p, q) &= \mathbb{E}_{Y \sim q}(p - Y)^2 \\ &= q(p-1)^2 + (1-q)p^2 \\ &= qp^2 - 2pq + q + p^2 - qp^2 \\ &= (p-q)^2 - q^2 + q \text{ (completing the square)}, \end{aligned}$$

  which is minimized by $p = q$.
- So this Brier score is proper.

Brier vs log-likelihood for evaluation?

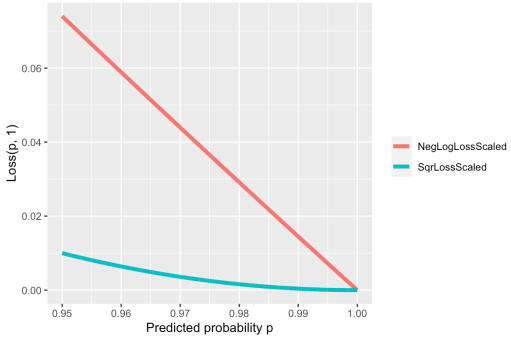
- Consider losses when true class is 1.
- Log-loss goes to infinity when we predict probability 0.
- Rescaling so they have the same value at $p = 0.5$:



Except for the far left, when we're very confidently predicting the wrong class, the two losses don't look dramatically different at this scale.

Brier vs log-likelihood for confident predictions

- With the same scaling as previous slide, let's zoom in



- Log-loss cares much more about getting the probability as close to 1.0 as possible

If we're in a domain where a probability of 0.990 is much different from 0.999, then perhaps log-loss makes more sense. Note that if these are success probabilities, then the corresponding failure rate for 0.99 is 10 times that of the failure rate for 0.999.

Conditional probability modeling for binary outcomes

Predicting binary outcome probabilities

- Suppose $(X, Y) \sim P_{\mathcal{X} \times \mathcal{Y}}$, where $Y \in \mathcal{Y} = \{0, 1\}$.
- Given observation $X = x$, we want to predict a distribution for Y.
- Consider prediction function $f : \mathcal{X} \to [0, 1]$
  - gives the predicted probability of $Y = 1$ for any $x \in \mathcal{X}$.
- The ideal evaluation of f is its expected score (i.e. its risk)

$$\mathbb{E}[S(f(X), Y)].$$

- We can estimate this empirically using a sample of data:

$$\frac{1}{n} \sum_{i=1}^{n} S(f(x_i), y_i).$$

Most common scores for binary outcomes: log-likelihood and square loss

Brier score / Mean squared error

$$\mathbb{E}[f(X) - Y]^2$$

and its empirical estimate

$$\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2.$$

Negative log-likelihood score

$$-\mathbb{E}\left[Y \log f(X) + (1-Y)\log(1-f(X))\right]$$

and its empirical estimate

$$-\sum_{i=1}^{n}\left[Y_i \log f(X_i) + (1-Y_i)\log(1-f(X_i))\right]$$

What minimizes the expected score for a binary outcome?

- With a proper score function, the minimizer of $\mathbb{E}[S(f(X), Y)]$ over all functions is

$$f(x) = \mathbb{P}(Y = 1 \mid X = x) = \mathbb{E}[Y \mid X = x].$$

- This follows simply by considering a single x at a time, and the "proper"-ness of S.
- In practice, your hypothesis space usually won't contain $\mathbb{E}[Y \mid X = x]$.
- Or even if it does, you'll probably need to regularize, essentially shrinking the hypothesis space.
- So in practice we don't actually find $\mathbb{E}[Y \mid X = x]$.

References

Resources

- The definitions used here for proper scoring rules are based on [GR07, Sec 4.1]

References I

[GR07] Tilmann Gneiting and Adrian E Raftery, *Strictly proper scoring rules, prediction, and estimation*, Journal of the American Statistical Association **102** (2007), no. 477, 359–378.