DATA 100: Vitamin 12 Solutions

November 8, 2019

1 L2 vs. Cross Entropy Loss

Which of the following are reasons to use cross entropy loss over L2 loss when computing the risk of a logistic regression? Select all that apply.

- ☑ The risk function computed using cross entropy loss is always convex.
- ☑ Cross entropy loss penalizes a wrong prediction more than L2 loss does.
- ☐ Cross entropy loss penalizes a wrong prediction less than L2 loss does.
- ☑ Cross entropy loss is more suited for comparing probability distributions than L2 loss.

Explanation: The first and fourth items are true, though we haven't proven these statements in class. You'll have to take our word for it! The second item is also true; a visual proof is provided on slide 29, lecture 20. This makes sense intuitively, since the L2 loss will be bounded between 0 and 1, whereas the log loss won't.

2 Computing Empirical Risk with Cross Entropy Loss

Suppose we fit a logistic regression model without an intercept term and find that $\hat{\beta} = [2, -3]^\top$. Compute the empirical risk the following dataset using cross entropy loss: $x_1 = [4, 0]^\top, y_1 = 0$ and $x_2 = [-1, -1]^\top, y_2 = 1$.

- ☐ $\frac{1}{2}[-\log(\sigma(1)) - \log(\sigma(8))]$
- ☑ $-\frac{1}{2}[\log(1 - \sigma(8)) + \log(\sigma(1))]$
- ☐ $-\frac{1}{2}[\log(\sigma(1)) + \log(1 - \sigma(1)) + \log(\sigma(8)) + \log(1 - \sigma(8))]$
- ☐ $\frac{1}{2}[-(1 - \log(\sigma(8))) - \log(\sigma(1))]$

Explanation: Recall that the empirical risk of a logisitc regression model is given by:

$$\begin{aligned} R(\beta) &= \frac{1}{n}\sum_{i=1}^{n} -\log\hat{\mathbb{P}}(Y = y_i|X = x_i) \\ &= -\frac{1}{n}\sum_{i=1}^{n} y_i \log\hat{\mathbb{P}}(Y = 1|X = x_i) + (1 - y_i)\log\hat{\mathbb{P}}(Y = 0|X = x_i) \end{aligned}$$

Where $\hat{\mathbb{P}}(Y = 1|X = x_i) = \sigma(x_i^\top \beta)$, and $\hat{\mathbb{P}}(Y = 0|X = x_i) = 1 - \sigma(x_i^\top \beta)$

Given that $n = 2$, $x_1^\top \beta = 8$, and $x_2^\top \beta = 1$, we find that:

$$R(\hat{\beta}) = -\frac{1}{2}[\log(1 - \sigma(8)) + \log(\sigma(1))]$$

3 Accuracy, Precision, and Recall

Suppose we've built a classifier to predict whether or not an image contains goats. We have 100 images, 65 of which contain goats.

- Our classifier predicted that 72 images contain goats. Of these, 52 truly contain goats, while the other 20 do not.
- Our classifier predicted 28 images to not contain goats. Of these, 13 truly contain goats, while the other 15 do not.

Sort the accuracy, precision, and recall of our classifier in increasing order.

☐ recall < accuracy < precision

☐ recall < precision < accuracy

☑ accuracy < precision < recall

☐ precision < recall < accuracy

Explanation: Recall the formulae for accuracy, precision and recall:

- Accuracy = $\frac{\text{\# of images correctly classified}}{\text{\# of images}}$
- Precision = $\frac{\text{\# of images with goats correctly classified}}{\text{\# of images with goats correctly classified+\# of images without goats classified as containing goats}}$
- Recall = $\frac{\text{\# of images with goats correctly classified}}{\text{\# of images with goats correctly classified+\# of images with goats classified as not containing goats}}$

And so we find that Accuracy = $\frac{52+15}{100} = 0.65$, Precision = $\frac{52}{52+20} \approx 0.722$ and Recall = $\frac{52}{52+13} = 0.8$.

4 Thresholding

As we increase our classification threshold, which of the following may happen?

- ☐ The number of false positives increases.
- ☑ The number of false positives decreases.
- ☑ The number of false negatives increases.
- ☐ The number of false negatives decreases.

Explanation: As the threshold for classification increases, the model assigns positive classifications much more stringently. Therefore, the number of positively observations classified observations decreases. In turn, this will likely decrease the number of false positives, but may also increase the number of false negatives. Stringking a balance between the two may be difficult; it is highly dependent of the area of application. For example, in a medical setting, we may tolerate high levels of false positives in order to reduce the number false negatives because of the "cost" of false negative (e.g. potentially the loss of human life) is much higher than the "cost" of a false postive (e.g. performing medical tests).

5 Imbalanced Datasets

Suppose we fit a classification model to predict binary outcome in a dataset where only 1% of the observations are positive, and the remaining are negative. Which of the following is not a reasonable metric to evaluate the model with?

- ☑ Accuracy
- ☐ Precision
- ☐ Recall

Explanation: Accuracy would be a poor metric to evaluate your model with since classifying all obersevations as negative would produce an accuracy level of 99%, which might mislead us to believe that we have produced a great model.