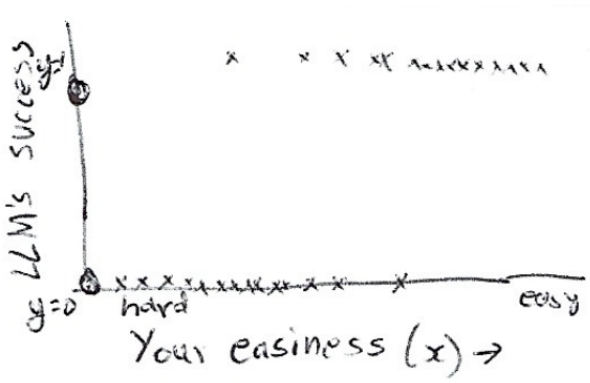
07-280 Logistic Regression

Nihar B. Shah

Suppose you are trying to evaluate an LLM. You ask it to solve a number of tasks. For each task, you check if it is successful or not. You then make the following plot:



$$\mathcal{X} = \mathbb{R} \quad \text{(your easiness, on some scale)}$$
$$\mathcal{Y} = \{0, 1\} \quad \text{(LLM successful or not)}$$

You may try to predict y from x.

Problem: for multiple tasks with same easiness, the LLM is sometimes successful and other times not. Even for the same task, under multiple runs, it is sometimes successful and at other times not. So, for instance, when it succeeds 50% of the time, it is not meaningful to predict y. All predictions will be wrong 50% of the time.

Hence, instead, we will try to estimate $\mathbb{P}(y = 1)$ for any x. This is now a regression problem since $\mathbb{P}(y = 1)$ is continuous valued.

What loss function to use? What model class to consider?
(e.g., 0–1 or squared etc.) (Previously: decision trees, linear)

Loss Function

Suppose you are predicting the probability of some event (i.e., whether it will occur). Your prediction of it happening was $\hat{p} \in [0, 1]$. *The event actually happened. Then what should the loss be?*

Some principles: If the event actually occurred, then a higher reported $\hat{p}$ should incur a lower loss. Moreover, the loss should capture the uncertainty expressed – a certain wrong prediction is worse than an uncertain wrong prediction. Recall from our earlier discsussions on entropy that

uncertainty is captured by the log of the reciprocal of the probability. So let's choose loss as $\log \frac{1}{\hat{p}}$.

More generally for our setting: $\ell : [0,1] \times \mathcal{Y} \to \mathbb{R}$. Under true label y and predicted probability $\hat{p}$, and when $\mathcal{Y} \in \{0,1\}$,

$$\ell(\hat{p}, y) = \mathbf{1}\{y=1\} \log \frac{1}{\hat{p}} + \mathbf{1}\{y=0\} \log \frac{1}{1-\hat{p}} = y \log \frac{1}{\hat{p}} + (1-y) \log \frac{1}{1-\hat{p}}. \qquad \text{"Cross entropy loss"}$$

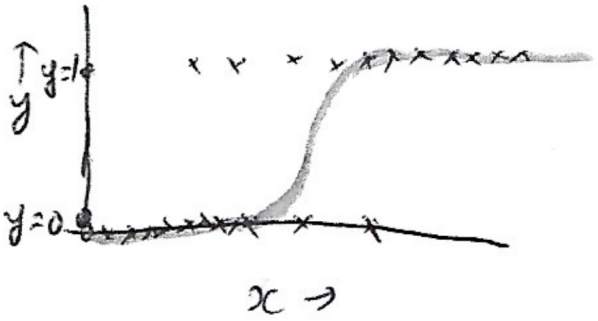
Model Class

Consider $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0,1\}$. Consider sigmoid (logistic) function $\sigma : \mathbb{R} \to [0,1]$:

$$\sigma(z) = \frac{1}{1+e^{-z}}.$$

Like in the case of (scalar) linear regression, our model is associated with two parameters $\omega \in \mathbb{R}$, $b \in \mathbb{R}$. Our logistic regression model assumes for any x:

$$\mathbb{P}(y=1) = \frac{1}{1+e^{-(\omega x + b)}}.$$



Under this model, given training data $\{(x^{(i)}, y^{(i)})\}_{i=1}^N$, we will compute ERM to find ω, b:

$$\arg\min_{\omega,b \in \mathbb{R}} \sum_{i=1}^N \ell(\hat{p}_\theta^{(i)}, y^{(i)}) = \arg\min_{\omega,b \in \mathbb{R}} \sum_{i=1}^N \left(y^{(i)} \log \frac{1}{\hat{p}_\theta^{(i)}} + (1-y^{(i)}) \log \frac{1}{1-\hat{p}_\theta^{(i)}} \right)$$

where

$$\hat{p}_\theta^{(i)} = \frac{1}{1+e^{-(\omega x^{(i)} + b)}}.$$

How to find the argmin in a computationally efficient manner? No closed form solution. We will come back to this later.

What about $\mathcal{X} = \mathbb{R}^d$?

$\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{0,1\}$. Consider parameter $\theta \in \mathbb{R}^d$ (and like in linear regression we assume that every x has a "1" prepended).

Model:

$$\mathbb{P}(y=1) = \frac{1}{1+e^{-\theta^\top x}}.$$

Then use the same expression for empirical risk minimization:

$$\arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^N \left(y^{(i)} \log \frac{1}{\hat{p}_\theta^{(i)}} + (1-y^{(i)}) \log \frac{1}{1-\hat{p}_\theta^{(i)}} \right), \quad \text{where } \hat{p}_\theta^{(i)} = \frac{1}{1+e^{-\theta^\top x^{(i)}}}.$$

Optimization

Can use GD or SGD. For any $i \in \{1, 2, \ldots, N\}$ let

$$J^{(i)}(\theta) = y^{(i)} \log \frac{1}{\hat{p}_\theta^{(i)}} + (1 - y^{(i)}) \log \frac{1}{1 - \hat{p}_\theta^{(i)}}, \quad \hat{p}_\theta^{(i)} = \frac{1}{1 + e^{-\theta^\top x^{(i)}}}.$$

Then

$$\nabla_\theta J^{(i)}(\theta) = \left(\hat{p}_\theta^{(i)} - y^{(i)}\right) x^{(i)} = \left(\frac{1}{1 + e^{-\theta^\top x^{(i)}}} - y^{(i)}\right) x^{(i)}.$$

Recall that linear regression ERM was a parabola, and hence there GD/SGD would come close to the optimum. But this is not a parabola. So now what?

Convexity

Consider function $f : \mathbb{R} \to \mathbb{R}$. Informally, the function is said to be convex if the line joining any two points is never below the function value. Formally, convex if

$$f(\lambda z_1 + (1 - \lambda) z_2) \leq \lambda f(z_1) + (1 - \lambda) f(z_2)$$

$$\forall\, z_1, z_2 \in \mathbb{R}, \quad \lambda \in [0, 1].$$

Another definition: $\dfrac{d^2}{dz^2} f(z) \geq 0 \quad \forall\, z \in \mathbb{R}$.

The aforementioned definitions also hold for functions $f : \mathbb{R}^d \to \mathbb{R}$ (although the second derivative needs to be defined carefully using multi-variate calculus).

Very useful property: This is also bowl-like. Every local minimum is also a global minimum. So use GD/SGD to find the optimum.

We can show that the expression for empirical risk under logistic regression is also convex in θ. As an exercise, try proving this for the scalar case $\mathcal{X} = \mathbb{R}$. Hence, GD/SGD can come very close to the actual minimum!

More than Two Classes

Suppose $\mathcal{Y} = \{1, 2, \ldots, K\}$. $\mathcal{X} = \mathbb{R}^d$. Then output is $\hat{p} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_K)$.

Cross-entropy loss: $\ell : [0, 1]^K \times \mathcal{Y} \to \mathbb{R}$

$$\ell(\hat{p}, y) = \sum_{j=1}^{K} \mathbf{1}\{y = j\} \log \frac{1}{\hat{p}_j}.$$

Model: has one weight vector associated to each class, i.e., $\theta_1, \ldots, \theta_K \in \mathbb{R}^d$.

$$\mathbb{P}(y = j) = \frac{e^{\theta_j^\top x}}{\sum_{\ell=1}^{K} e^{\theta_\ell^\top x}} \quad \text{for any } x \in \mathbb{R}^d.$$

This is called "softmax regression" or "multinomial logistic regression." One then performs ERM under this loss and this model class.