Runqiu Ye
Stanford CS299
Problem Set #2
06/30/2024

Problem Set #2: Supervised Learning II

Problem 1 Logistic Regression: Training stability

(a) The most notable difference in training the logistic regression model on datasets A and B is that the training process on dataset B requires far more iterations to converge.

(b)

$\blacksquare$

Problem 2 Model Calibration

Try to understand the output $h_\theta(x)$ of the hypothesis function of a logistic regression model, in particular why we might treat the output as a probability.

When probabilities outputted by a model match empirical observation, the model is *well-calibrated*. For example, if a set of examples $x^{(i)}$ for which $h_\theta(x^{(i)}) \approx 0.7$, around 70% of those examples should have positive labels. In a well-calibrated model, this property holds true at every probability value.

Suppose training set $\{x^{(i)}, y^{(i)}\}_{i=1}^m$ with $x^{(i)} \in \mathbb{R}^{n+1}$ and $y^{(i)} \in \{0, 1\}$. Assume we have an intercept term $x_0^{(i)} = 1$ for all i. Let θ be the maximum likelihood parameters learned after training logistic regression model. In order for model to be well-calibrated, given any range of probabilities (a, b) such that $0 \leq a < b \leq 1$, and trianing examples $x^{(i)}$ where the model outpus $h_\theta(x^{(i)})$ fall in the range (a, b), the fraction of positives in that set of examples should be equal to the average of the model outputs for those examples. That is,

$$\frac{\sum_{i \in I_{a,b}} P(y^{(i)} = 1 \mid x^{(i)}; \theta)}{|\{i \in I_{a,b}\}|} = \frac{\sum_{i \in I_{a,b}} \mathbf{1}\{y^{(i)} = 1\}}{|\{i \in I_{a,b}\}|},$$

where $P(y^{(i)} = 1 \mid x; \theta) = h_\theta(x) = 1/(1 + \exp(-\theta^T x))$, $I_{a,b} = \{i : h_\theta(x^{(i)}) \in (a, b)\}$.

(a) For the described logistic regression model over the range $(a, b) = (0, 1)$, we want to show the above equality holds. Recall the gradient of log-likelihood

$$\frac{\partial \ell}{\partial \theta_j} = \sum_{i=1}^m (y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)}.$$

For a maximum likelihood estimation, $\frac{\partial \ell}{\partial \theta} = 0$. Hence $\frac{\partial \ell}{\partial \theta_0} = 0$. Since $x_0^{(i)} = 1$, we have

$$\sum_{i=1}^{m} y^{(i)} - h_\theta(x^{(i)}) = 0.$$

The desired equality follows immediately.

(b) A perfectly calibrated model — that is, the equality holds for any $(a, b) \subset [0, 1]$ — does not imply that the model achieves perfect accuracy. Consider $(a, b) = (\frac{1}{2}, 1)$, the above equality implies

$$\frac{\sum_{i \in I_{a,b}} P(y^{(i)} = 1 \mid x^{(i)}; \theta)}{|\{i \in I_{a,b}\}|} = \frac{\sum_{i \in I_{a,b}} \mathbf{1}\{y^{(i)} = 1\}}{|\{i \in I_{a,b}\}|} < 1.$$

This shows that the model does not have perfect accuracy.

For the converse direction, a perfect accuracy does not imply perfectly calibrated. Consider again $(a, b) = (\frac{1}{2}, 1)$, then we have

$$\frac{\sum_{i \in I_{a,b}} \mathbf{1}\{y^{(i)} = 1\}}{|\{i \in I_{a,b}\}|} = 1 > \frac{\sum_{i \in I_{a,b}} P(y^{(i)} = 1 \mid x^{(i)}; \theta)}{|\{i \in I_{a,b}\}|}.$$

(c) Discuss what effect of L_2 regularization in the logistic regression objective has on model calibration.

■

$(0, 1)$ is the only range for which logistic regression is guaranteed to be calibrated. When GLM assumptions hold, all ranges $(a, b) \subset [0, 1]$ are well calibrated. In addition, when test set has same distribution and when model has not overfit or underfit, logistic regression are well-calibrated on test data as well. Thus logistic regression is popular when we are interested in level of uncertainty in the model output. △