Machine Learning 4002: Assignment #1

Due on Farvardin 17, 1401

Dr. Mehran Safayani Fall 1400

Arian Tashakkor

40023494

Problem 1

A.

Statement is incorrect. Training accuracy cannot be used as a gauge to measure how well a model performs as the performance of a model on the training data does not necessarily extend to its performance on the test data. For instance, take the case of "overfitting". Imagine a model with billions of parameters designed to solve a relatively simple task. This model, given enough time in epochs, will eventually achieve a near perfect accuracy on the training data - because we know that a sophisticated enough neural network is able to imitate any function. However, this model will inevitably overfit on the training data and cannot generalize well, meaning that it is not necessarily "better" by any measure simply because it performs better on the training dataset.

B.

Statement is incorrect. Because the positive instances have extremely low support value against the negative instances, even if the model is trained on the 80% of the data containing only negative instances (that is 40040 samples of negatives), assuming that it generalizes well enough to be able to categorize all the negative instances in the test data, even if it misclassifies every positive instance as a False Negative, it will still have an error margin of about 0.5% (or an accuracy of about 99.5%). As such, in order to evaluate a model, Precision, Recall and F1-Score that combines both metrics are much better measures.

C.

Statement is correct. This is indeed the scientifically correct way of conducting a research that employs a learning technique. If the training, validation and test splits of the dataset are proportionally consistent, we can hope to achieve about the same accuracy on the test dataset as we do on the validation dataset. Therefore, it makes sense to choose the model that performs best on the validation dataset and report its accuracy/error on the test dataset.

D.

Statement is partially incorrect. It is a good practice to perform dimensionality reduction on extremely multifaceted datasets. However, reporting accuracy/error directly on the test dataset is not scientifically correct as the model should never, outside of deployment, see the test dataset. Therefore, tuning the hyperparameters of the model to achieve the best accuracy on the test dataset is essentially akin to overfitting the model on that dataset and thereby removes all significance of the reported final accuracy because the reported metrics should be trusted to be that model's performance on data it has never seen before and be a measure of how well it can generalize. A purpose which is defeated entirely by reporting the best accuracy on the test dataset.

Problem 2

We will first define each of the cost functions employed as follows:

- NR (No Regularization):

$$\min_w \sum_{i=1}^{n} (y_i - w^T x_i)^2$$

- L2R (L2 Regularization):

$$\min_w \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda \sum_{j=1}^{8} {w_j}^2$$

- L1R (L1 Regularization):

$$\min_w \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda \sum_{j=1}^{8} |w_j|$$

We know that regularization is a process that helps keep the magnitude of a model's weights in check; enforcing them to be as small as possible or otherwise suffer a penalty in the resulting loss which is back-propagated through the model. Therefore, we expect the weights of a model that uses NR to be larger in absolute value than the same model when it uses some form of regularization. Additionally, L2R attempts to keep the weights close to 0, but it cannot completely zero out a weight unless λ tends to infinity[1]. Therefore, if no manual feature selection has been conducted and the trained weights of a model contain zero values, it was very likely done using L1R. Now we can begin with the observations that help us figure out which regularization belongs to which column of weights:

- Column A appears to be larger in absolute value than the other columns. Indicating that it was likely the trained weights of a model that used NR.
- Column C has zeroed-out weights, which is only possible when LASSO regularization is used, indicating that this column is likely the trained weights of a model that used L1R.
- That leaves Column B which is, by process of elimination, likely the trained weights of a model that used L2R. Note that this column has weights close to 0 but not exactly 0.

[1] `https://towardsdatascience.com/l1-and-l2-regularization-explained-874c3b03f668`

Problem 3

Assuming the samples were drawn independently from one another, the joint likelihood of all of these samples can be calculated by the product of their individual probabilities. There are:

- 3 observations of "1" at $P(X=1)=\frac{\theta}{3}$.
- 3 observations of "2" at $P(X=2)=\frac{2(1-\theta)}{3}$.
- 2 observations of "0" at $P(X=0)=\frac{2\theta}{3}$.
- And 2 observations of "3" at $P(X=3)=\frac{1-\theta}{3}$.

The joint likelihood, $J(\theta)$, is then calculated as:

$$J(\theta)=\prod_{s\in Samples} P(s)=(\frac{\theta}{3})^3(\frac{2(1-\theta)}{3})^3(\frac{1-\theta}{3})^2(\frac{2\theta}{3})^2$$

Since ln is a monotonically increasing function, the extrema of $J(\theta)$ and $\ln(J(\theta))$ coincide, and it is easier to differentiate $\ln(J(\theta))$ because it turns the product term into a sum of components, each of which can be individually differentiated. Therefore, we have:

$$\begin{aligned}\ln(J(\theta)) &= 3(\ln\frac{\theta}{3})+3(\ln\frac{2(1-\theta)}{3})+2(\ln\frac{1-\theta}{3})+2(\ln\frac{2\theta}{3})\\ &= 3(\ln\theta-\ln 3)+3(\ln 2+\ln(1-\theta)-\ln 3)+2(\ln(1-\theta)-\ln 3)+2(\ln 2+\ln\theta-\ln 3)\\ &= 5\ln\theta+5\ln(1-\theta)+C,\ \text{where C is some constant.}\end{aligned}$$

Differentiating w.r.t. θ and setting to 0 we have:

$$\begin{aligned}\frac{\partial}{\partial\theta}(\ln(J(\theta))=\frac{5}{\theta}-\frac{5}{1-\theta}=5(\frac{1}{\theta}-\frac{1}{1-\theta})=0\\ \rightarrow\frac{1-2\theta}{\theta-\theta^2}=0\rightarrow\theta=\frac{1}{2}\end{aligned}$$

Problem 4

Similar to the previous problem, since the samples are said to be i.i.d, the joint likelihood is the product of the individual marginal probabilities. I.e.:

$$J(\sigma) = \prod_{i=1}^{n} P(x_i;\sigma) = \prod_{i=1}^{n} \frac{1}{2\sigma} e^{-\frac{|x_i|}{\sigma}} = (\frac{1}{2\sigma})^n e^{-\frac{1}{\sigma}\sum_{i=1}^{n}|x_i|}$$

Again we will work with the $\ln J(\sigma)$ for ease of differentiation:

$$\ln J(\sigma) = n(\ln 1 - \ln 2 - \ln \sigma) - \frac{1}{\sigma}\sum_{i=1}^{n}|x_i| =$$

$$-n\ln\sigma - \frac{1}{\sigma}A + C, \text{ where C is some constant and A} = \sum_{i=1}^{n}|x_i|$$

Differentiating w.r.t. σ and setting to 0 we have:

$$\frac{\partial}{\partial\sigma}(\ln J(\sigma)) = -\frac{n}{\sigma} + \frac{A}{\sigma^2}$$

$$\rightarrow -\frac{n}{\sigma} + \frac{A}{\sigma^2} = 0 \rightarrow \frac{A - n\sigma}{\sigma^2} = 0$$

$$\rightarrow \sigma = \frac{A}{n} = \frac{1}{n}\sum_{i=1}^{n}|x_i|$$

Observation

It seems that the distribution provided in this problem is the one that stems from using a LASSO regression in a cost function.