Introduction to Machine Learning

Nizar Ghandri
Home Assignment 2
ENS ULM

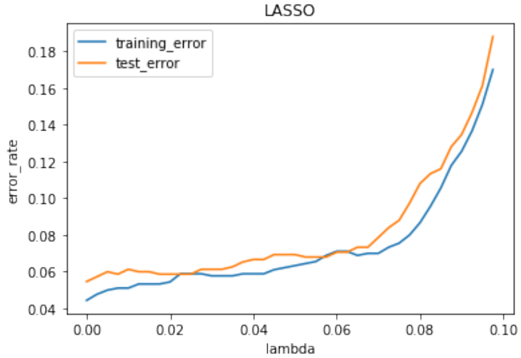
May 21, 2020

1 Regularization of logistic regression

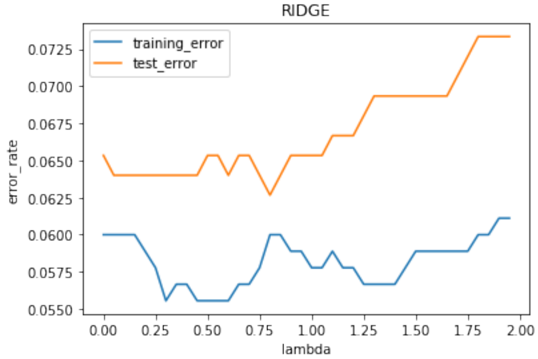
1.Why is it often important to regularize?

Regularizing prevents us from overfitting the model on the training data. Adding the penalty term $||\beta||_1$ for LASSO or $||\beta||_2^2$ for Ridge allows us to actually reduce the variance but having a larger bias for our estimator.

To minimize LASSO we implement the ISTA algorithm, (for the proximal operator of the l1 norm, it's the soft-thresholding operator as mentioned in the course). Where as the minimize ridge we use a simple gradient descent given the function is differentiable everywhere.

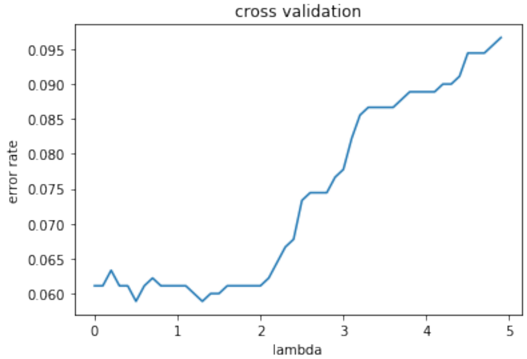


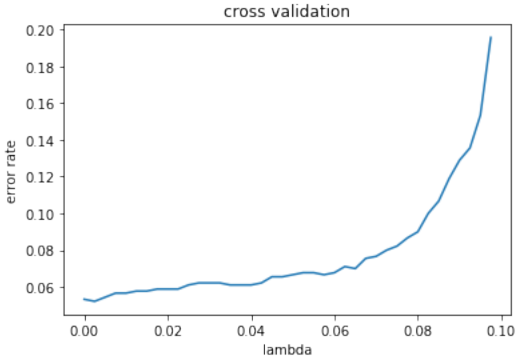
we can see how around λ =0.02 and 0.06 we lose training precision to gain in test precision.



we can see how around λ =0.8 we lose training precision to gain in test precision. 3.What would be the best value for λ ? How would you tune it?

Ridge: 0.5
LASSO: 0.0025
We find these values with a K-fold cross validation:





2 LDA

a) how that the conditional model of $P(Y = 1|X)$ associated with this generative model is of the form

$$P(Y = 1|X) = \frac{1}{1 + e^{(-beta_0 - <\beta, X>}}$$

,for som e$\beta_0 \in \mathbb{R}$ and $\beta \in \mathbb{R}$ depending on $\sum, \mu_{-1}, \mu_1$,and π to be explicited.

$$\begin{aligned}
P(Y = 1|X)n &= \frac{P(X|Y = 1) * P(Y = 1)}{P(X|Y = 1) * P(Y = 1) + P(X|Y = -1) * P(Y = -1)} && (1) \\
&= \frac{\frac{\pi}{(2\pi)^{\frac{d}{2}} |\sum^{-1}|^{\frac{1}{2}}} e^{(x-\mu_1)^T \sum (x-\mu_1)}}{\frac{\pi}{(2\pi)^{\frac{d}{2}} |\sum^{-1}|^{\frac{1}{2}}} e^{(x-\mu_1)^T \sum (x-\mu_1)} + \frac{1-\pi}{(2\pi)^{\frac{d}{2}} |\sum^{-1}|^{\frac{1}{2}}} e^{(x-\mu_{-1})^T \sum (x-\mu_{-1})}} && (2) \\
&= \frac{1}{1 + \frac{\pi e^{(x-\mu_1)^T \sum (x-\mu_1)}}{(1-\pi) e^{(x-\mu_{-1})^T \sum (x-\mu_{-1})}}} && (3) \\
&= \frac{1}{e^{0.5 \log(\frac{1-\pi}{\pi})(\mu_{-1}^T \sum^{-1} \mu_{-1} + \mu_1^T \sum^{-1} \mu_1 + 2x^T \sum^{-1} (\mu_1 - \mu_{-1})}} && (4)
\end{aligned}$$

Thus $\beta_0 = 0.5 \log(\frac{1-\pi}{\pi})(\mu_{-1}^T \sum^{-1} \mu_{-1} + \mu_1^T \sum^{-1} \mu_1)$ and $\beta_1 = 0.5 \log(\frac{1-\pi}{\pi}) 2x^T \sum^{-1} (\mu_1 - \mu_{-1})$
b) Show that, assuming $\beta_0 = 0$(i.e., no intercept), the maximum likelihood of β of this probabilistic model corresponds to the solution of logistic regression with $\lambda = 0$.

$$\begin{aligned}
P(Y = -1|X) &= 1 - P(Y = 1|X) && (5) \\
&= \frac{e^{-\beta^T X}}{1 + e^{-\beta^T X}} && (6) \\
&= \frac{1}{1 + e^{\beta^T X}} && (7)
\end{aligned}$$

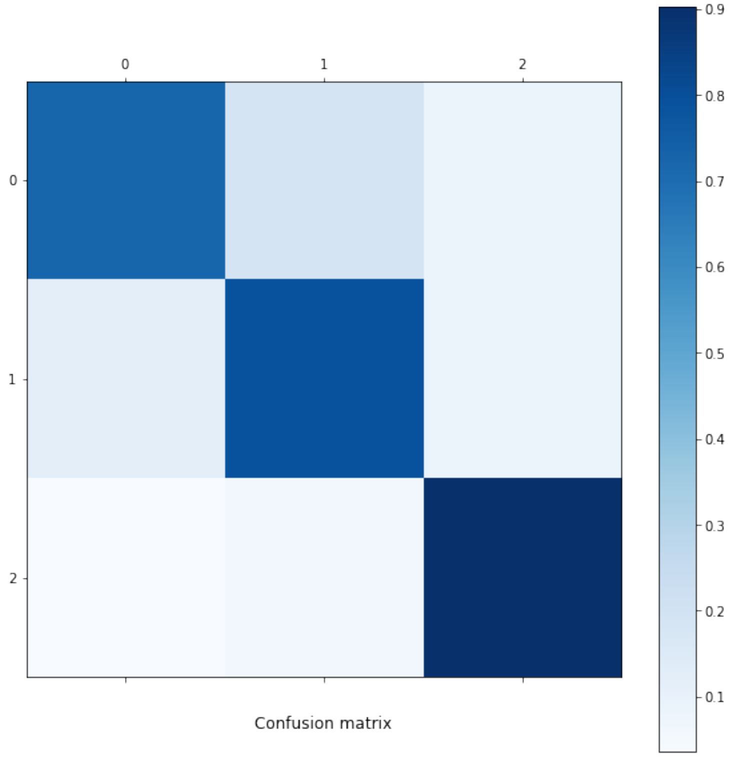
Thus

$$P(Y = y|X) = \frac{1}{1 + e^{-y\beta^T X}}$$

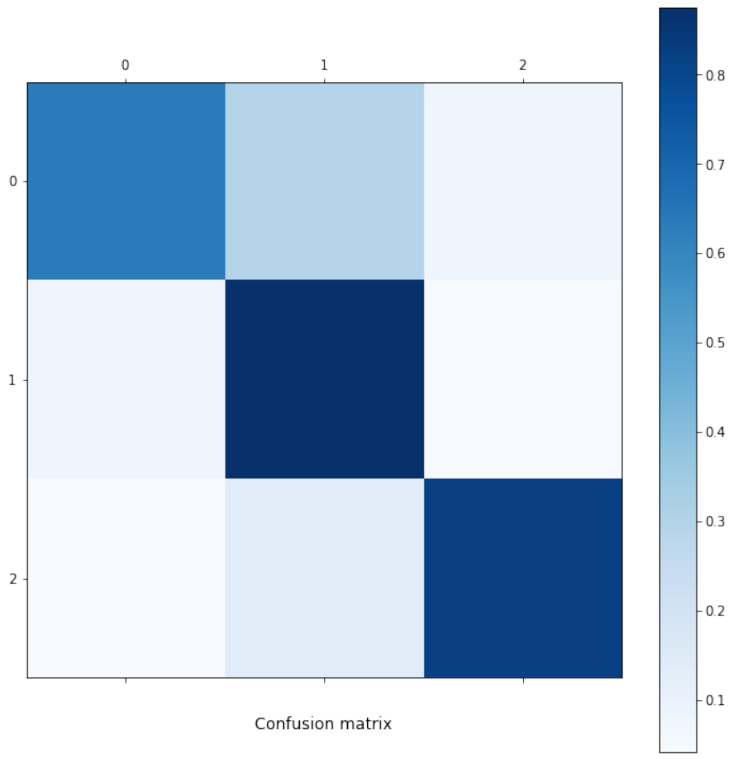
c) couldn't solve numerical instability ofr this problem d)In our case false negative is when we have predicted a -1 but the real value was a 1, a false positive is when have predicted a 1 but the real value is -1. A confusion matrix shows all four values (false positive, false negative, true positive and true negative) and is more relevant fr clustering algorithms compared to classification error. It's also more relevant in unbalanced data sets. It allows you to see how homogeneous some clusters are compared to others thus giving more insight than classification error.

3 K means

K means although very sensitive to its initialization ends up, if we always select the best our 10 executions, classifying each cluster to a class with an average of 0.82 of purity for the three clusters:



To try and get better results I reasoned that PCA should be done before to avoid the dimensionality curse (l2 becomes less less relevant in higher dimension spaces) however 2 components represented very few features that were not enough for a good separation, the model acheived a purity of 0.78 for the three clusters:


0
1
2
0
1
2
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Confusion matrix