Machine Learning for Computer Vision

[MVA 2015/2016]

Programming Assignment 1

Maha ELBAYAD

1 Linear versus logistic regression

Our objective is to maximize the criterion C on the training set $\mathcal{X} = (x_i, y_i)_{1 \leq i \leq N}$ ($x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$)

$$C(w) = \sum_{i=1}^{N} y_i \log\left(g\langle x_i, w\rangle\right) + (1 - y_i)\log\left(1 - g\langle x_i, w\rangle\right)$$

where g is the sigmoid function $g(x) = \frac{1}{1+\exp(-x)}$.
We can rewrite $C(w)$ as:

$$C(w) = \sum_{i=1}^{N} \log(1 - g\langle x_i, w\rangle) + y_i\langle x_i, w\rangle$$

Knowing that $g'(a) = g(a)(1 - g(a)) \, \forall a \in \mathbb{R}$

$$\nabla_w C(w) = J(w) = \sum_{i=1}^{N} (y_i - g\langle x_i, w\rangle) x_i$$

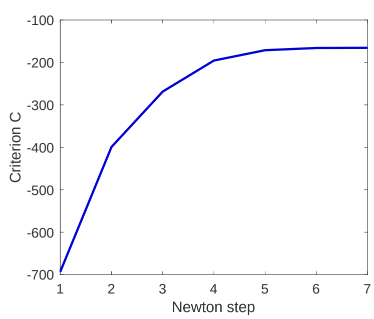
$$\nabla_w^2 C(w) = H(w) = \sum_{i=1}^{N} -x_i x_i^T g(\langle x_i, w\rangle)(1 - g(\langle x_i, w\rangle)$$

if we introduce the design matrix $X = [x_1, ...x_N] \in \mathbb{R}^{N \times d}$, the probabilities vector $G_w = [g\langle x_i, w\rangle]_i$ and $Y = [y_1, ..y_N]^T$ we can deduce a matrix form for both J and H:

$$J(w) = X^T(Y - G_w)$$
$$H(w) = -X^T D X, \; D = diag(G_w \odot (1 - G_w))$$

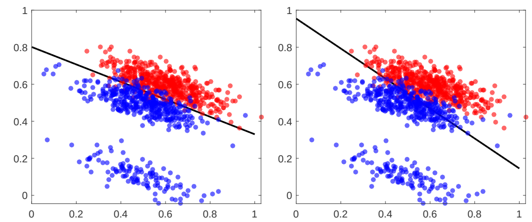
The Newton Raphson update would be:

$$w = w_{prev} - H(w_{prev})^{-1} J(w_{prev})$$



Newton-Raphson convergence - logistic regression

The linear regression is sensitive to the presence of outliers while the logistic regression seems more robust which leads to better accuracy



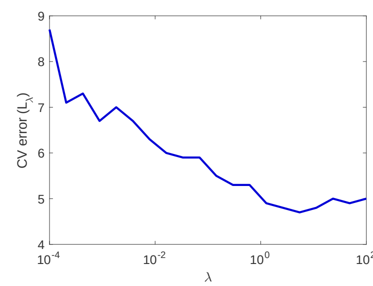
(a) *Logistic regression error rate = 6.40%*
(b) *Linear regression error rate = 11.20%*

2 Logistic Regression and Regularization

With the additional l_2 regularization term:

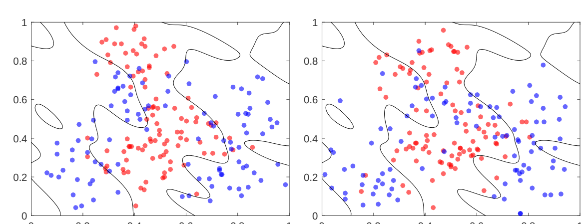
$$J(w) = X^T(Y - G_w) - 2\lambda w$$
$$H(w) = -X^T D X - 2I_d$$

with a K-fold cross validation we estimate the most appropriate λ from a set of values ranging from 1e-4 to 100:



K-fold average error as a function of λ
best performance at $\lambda^* = 5.4556$

With the logistic regression applied on a training set of size 200 with embedding dimension 241 (plus the bias term) the model overfits the data with fitted probabilities that are too close to 0/1 (an infinite criterions $C(w)$) and performs poorly on the test set.



(c) *Training set Error rate 14.5%*
(d) *Test set Error rate 24.50%*

Decision boundaries $\lambda = \lambda^*$