Maura Pintor, PhD Student

 maurapintor

 https://maurapintor.github.io

 @maurapintor

 https://www.linkedin.com/in/maura-pintor

 maura.pintor@unica.it

- Given a regression problem and the following data samples

$$\mathbf{X} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ 0 & 0 \\ 0 & 2 \\ 1 & 2 \\ 2 & 1 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \end{bmatrix},$$

- find the linear discriminant function via ordinary least squares (OLS), i.e., by minimizing:

$$L_r(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}^T\boldsymbol{x}_i + b - y_i)^2$$

- Initialize $\boldsymbol{w} = [0.1, 0.1]^T,\ b = 0.1,\ \eta = 0.1,\ \theta = 0.06$

How is this computed?

iter	$f(\boldsymbol{x})$	L_r	$\nabla_{\boldsymbol{w}} L_r, \nabla_b L_r$	$\boldsymbol{w}, b$	t
1.	0.1*[0 0 1 3 4 4]	0.240	[1. 1.4], 0.6	[0. -0.04], 0.04	0.3
2.	0.01*[4 8 4 -4 -4 0]	0.057	[-0.28 -0.64], -0.52	[0.028 0.024], 0.092	0.14
3.	0.01*[6 7 9 14 17 17]	0.014	[0.25 0.32], 0.104	[0.0032 -0.008], 0.0816	0.07
4.	0.01*[8 9 8 7 7 8]	0.003	[-0.05 -0.14], -0.136	[0.008 0.006], 0.09	0.03

θ=0.06

Exercise 1 (13 points)
Given the n=6 two-dimensional data points **x**, and their labels **y**

$$\mathbf{x} = \begin{bmatrix} -2 & 0 \\ 0 & -2 \\ -1 & -1 \\ 1 & 2 \\ 1 & 1 \\ 2 & 0 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \end{bmatrix},$$

find the linear discriminant function using a **batch gradient-descent algorithm** to minimize the following objective function (using the so-called *ramp loss*):

$L(\boldsymbol{w}, b) = \frac{1}{n}\sum_{i=1}^{n} [\max(0,1 - y_i f(\boldsymbol{x}_i)) - \max(0, -y_i f(\boldsymbol{x}_i))] + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}$, where $f(\boldsymbol{x}_i) = \boldsymbol{w}^T\boldsymbol{x}_i + b$.

The gradient of $l(yf(x)) = [\max(0,1 - yf(\boldsymbol{x})) - \max(0, -yf(\boldsymbol{x}))]$ with respect to the classifier parameters is not null only when $0 < yf(\boldsymbol{x}) < 1$, and in that case it equals $-y\,\boldsymbol{x}$ and $-y$, respectively when computed with respect to $\boldsymbol{w}$ and b.

Initialize $\boldsymbol{w} = [0.2, \ -1]^T$, $b = 0$, the gradient step size $\eta = 0.5$, the parameter $\lambda = 0.5$, and the threshold on the termination condition $\theta = 0.7$.

Use the l1 norm to compute $|\nabla_w L \ (\boldsymbol{w}, b)| + |\nabla_b L \ (\boldsymbol{w}, b)|$ in the termination condition.

- State the gradient-descent learning algorithm.
- Compute $\boldsymbol{w}, b$ for the first two iterations of the algorithm, and check if it converges.
- Plot the initial decision boundary along with the training points, and how it changes during the first two iterations of the algorithm.
- Plot the ramp loss function $l(z) = [\max(0,1 - z) - \max(0, -z)]$, with respect to z. Explain why the gradient is not null only when $0 < z < 1$.
- The ramp loss is more robust to outliers in the training data. Can you explain why?

https://github.com/unica-ml/ml/blob/master/exams/20180531-pre2a.pdf

$$\nabla_w L = \begin{cases} \frac{1}{n}\sum(-y_i \vec{x}_i) & \text{if } (y_i f(x_i) \text{ in } (0,1)) \\ 0 & \text{otherwise} \end{cases} + \lambda \vec{w}$$

$$\nabla_b L = \begin{cases} \frac{1}{n}\sum -y_i & \text{if } f(x_i) \cdot y_i \text{ is in } (0,1) \\ 0 & \text{otherwise} \end{cases} + 0$$

Exercise 2 (12 points)

Given the two-dimensional data points $\mathbf{x}$ in Exercise 1, and the initial k=2 centroids $\mathbf{v} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$:

$$\mathbf{x} = \begin{bmatrix} -2 & 0 \\ 0 & -2 \\ -2 & -1 \\ 1 & 1 \\ 1 & 0 \\ 2 & 1 \end{bmatrix}$$

- Cluster the data points x using the k-means algorithm, reporting the clustering labels, the updated centroids and the objective function at each iteration of the algorithm. For simplicity, use the L1 (Manhattan) distance instead of the L2 (Euclidean) distance, both for computing the objective function $\sum_{i=1}^{n} \left|\left|\boldsymbol{x}_i - \boldsymbol{v}_k\right|\right|_1$ (being $\boldsymbol{v}_k$ the closest centroid to $\boldsymbol{x}_i$) and for computing the distances between the data points x and the centroids v. If a point has the same distance with respect to a number of centroids, assign it to the centroid with the lowest class index in this set (e.g., if the point has the same distance w.r.t. centroid 0 and 2, assign it to centroid 0).
- Make a two-dimensional plot displaying the data points (with a clear indication to explain to which cluster each point belongs to, after the last iteration) and the final centroids.
- Plot the decision boundaries of the nearest mean centroid classifier that uses the final centroids of the k-means algorithm as the estimated centroids of each class.

https://github.com/unica-ml/ml/blob/master/exams/20190604-pre2.pdf

Exercise 3 (11 points)

Let us consider a 3-class problem in R^2 (two-dimensional feature space), where the likelihood of each class is Gaussian and given as $p(x|\omega_i) = N(\boldsymbol{\mu_i}, \Sigma_i)$, with

$\Sigma_i = \sigma^2\mathbf{I}$; $\mu 1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$; $\mu 2 = \begin{pmatrix} +1 \\ -1 \end{pmatrix}$; $\mu 3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$;

and prior probabilities $P_1 = P_2 = P_3$.

- Compute the decision boundaries and plot them.

https://github.com/unica-ml/ml/blob/master/exams/20190415-pre1.pdf

Links used today

- MachineLearningCheatSheet.pdf
- ML-tutor-04-whiteboard