Homework 2 – 孟念 Niklas Muennighoff

AdaBoost

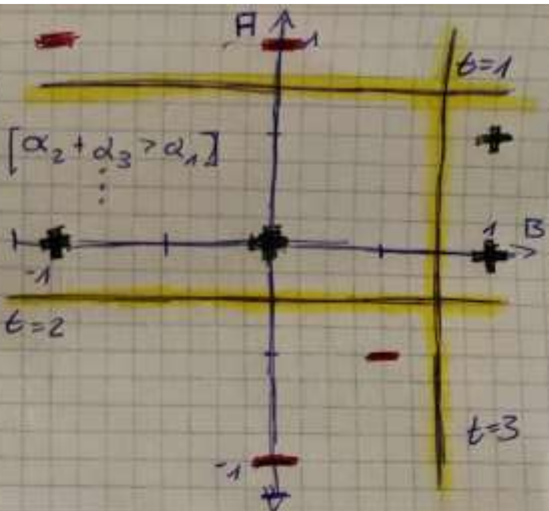
a)

ADABOOST

A	B	y	t_1	t_2	t_3	$\hat{y}$
-1	0	−	+	−	−	−
-0.5	0.5	−	+	−	−	−
0	1	+	+	+	+	+
0.5	1	+	+	+	+	+
1	0	−	−	+	−	−
1	-1	−	−	+	−	−
0	-1	+	+	+	−	+
0	0	+	+	+	−	+

$[\alpha_2 + \alpha_3 > \alpha_1]$

⋮



$t=1$: $A < 0.75$

True / \ False

+ −

$\varepsilon_1 = \frac{2}{8}$ $\alpha_1 = \frac{1}{2}\ln\left(\frac{1-\frac{2}{8}}{\frac{2}{8}}\right) = \frac{1}{2}\ln(3)$

$D_1^{\text{incorrect}} = \frac{1}{8} \cdot e^{\frac{1}{2}\ln(3)} = \frac{1}{8} \cdot \sqrt{3}$; $D_1^{\text{correct}} = \frac{1}{8 \cdot \sqrt{3}} \rightarrow Z_1 = \frac{\sqrt{3}}{8} \cdot 2 + \frac{1}{8\sqrt{3}} \cdot 6 = \frac{2 \cdot 3}{8\sqrt{3}} + \frac{6}{8\sqrt{3}}$

$\xrightarrow{\text{Norm}} D_1^{\text{incorrect}} = \frac{\sqrt{3}}{8} : \frac{3}{2\sqrt{3}} = \frac{6}{24} = \frac{3}{12}$; $D_1^{\text{correct}} = \frac{1}{12}$ $= \frac{12}{8\sqrt{3}} = \frac{3}{2\sqrt{3}}$

$t=2$: $A > -0.25$ $\varepsilon_2 = 2/12$; $\alpha_2 = \frac{1}{2}\ln\left(\frac{1-\frac{2}{12}}{\frac{2}{12}}\right) = \frac{1}{2}\ln(5)$

True / \ False

+ −

$D_2^{\text{correct}\rightarrow\text{incorrect}} = \frac{1}{12} \cdot e^{\frac{1}{2}\ln(5)} = \frac{\sqrt{5}}{12}$; $D_2^{\text{incorrect}\rightarrow\text{correct}} = \frac{3}{12} e^{-\frac{1}{2}\ln(5)} = \frac{3}{12\sqrt{5}}$

$D_2^{\text{correct}\rightarrow\text{correct}} = \frac{1}{12} e^{-\frac{1}{2}\ln(5)} = \frac{1}{12\sqrt{5}}$

$Z_2 = \frac{\sqrt{5}}{12} \cdot 2 + \frac{3}{12\sqrt{5}} \cdot 2 + \frac{1}{12\sqrt{5}} \cdot 4 = \frac{2 \cdot 5 + 6 + 4}{12\sqrt{5}} = \frac{5}{3\sqrt{5}}$

$\xrightarrow{\text{Norm}} D_2^{\text{correct}\rightarrow\text{incorrect}} = \frac{\sqrt{5}}{12} \cdot \frac{3\sqrt{5}}{5} = \frac{3}{12} = \frac{5}{20}$ $D_2^{\text{incorrect}\rightarrow\text{correct}} = \frac{3}{12\sqrt{5}} \cdot \frac{3\sqrt{5}}{5} = \frac{9}{60} = \frac{3}{20}$

$D_2^{\text{correct}\rightarrow\text{correct}} = \frac{1}{12\sqrt{5}} \cdot \frac{3\sqrt{5}}{5} = \frac{1}{20}$

$t=3$: $B > 0.75$ $\varepsilon_3 = \frac{2}{20}$ $\alpha_3 = \frac{1}{2}\ln\left(\frac{1-\frac{2}{20}}{\frac{2}{20}}\right) = \frac{1}{2}\ln(4)$

True / \ False

+ −

$D_3^{\text{correct}\rightarrow\text{correct}\rightarrow\text{correct}} = \frac{1}{20\sqrt{9}}$

$D_3^{\text{correct}\rightarrow\text{correct}\rightarrow\text{incorrect}} = 1/20 \cdot e^{\frac{1}{2}\ln(9)} = \frac{\sqrt{9}}{20}$ $D_3^{\text{correct}\rightarrow\text{incorrect}\rightarrow\text{correct}} = \frac{5}{20\sqrt{9}}$

$D_3^{\text{incorrect}\rightarrow\text{correct}\rightarrow\text{correct}} = \frac{3}{20\sqrt{9}}$

$Z_3 = \frac{\sqrt{9}}{20} \cdot 2 + \frac{1}{20\sqrt{9}} \cdot 2 + \frac{5}{20\sqrt{9}} \cdot 2 + \frac{3}{20\sqrt{9}} \cdot 2 = \frac{36}{20\sqrt{9}} = \frac{9}{5\sqrt{9}}$

$\xrightarrow{\text{Norm}} D_3^{\text{correct}\rightarrow\text{correct}\rightarrow\text{incorrect}} = \frac{\sqrt{9}}{20} \cdot \frac{5\sqrt{9}}{9} = \frac{5}{20}$ $D_3^{\text{correct}\rightarrow\text{correct}\rightarrow\text{correct}} = \frac{1}{20\sqrt{9}} \cdot \frac{5\sqrt{9}}{9} = 1/36$

$D_3^{\text{correct}\rightarrow\text{incorrect}\rightarrow\text{correct}} = 5/36$

$D_3^{\text{incorrect}\rightarrow\text{correct}\rightarrow\text{correct}} = 3/36$

b)

The y_hat column in a) shows that the trained classifier correctly predicts all training examples hence its training error is 0.

AdaBoost outperforms a single decision stump, because it combines multiple stumps that correct the previous stumps mistakes. All stumps in AdaBoost are then weighted at prediction.

AdaBoost is hence guaranteed to be at least as good as a single stump.

K – Means Clustering

a)

No we cannot minimize the function over both K and c at the same time.

To choose a good K, we can compute the total error for each K and then plot the different K’s with their sum. If the result is an ‘elbow plot’, we choose the elbow as our optimal K value.

b)

b)

1) Let j be the j^{th} cluster $\in \{1, 2 \ldots k\}$

Let i be the i^{th} sample $\in \{1, 2 \ldots n\}$

Let $z_{ij} = \begin{cases} 1 & \text{, if } x^{(i)} \text{ is assigned to } c_j \\ 0 & \text{else} \end{cases}$ (sample i assigned to cluster j)

$\rightarrow \sum_{i=1}^{n} z_{ij}$ corresponds to how many samples in cluster j

Assign to closest cluster:

$$\rightarrow z'_{ij} = \begin{cases} 1, & \|x^{(i)} - c_j\|_2^2 \leq \|x^{(i)} - c_{j'}\|_2^2, \forall j' \\ 0, & \text{else} \end{cases}$$

Normalize:

$$\rightarrow \alpha_{ij} = z'_{ij} \cdot \frac{1}{\sum_{i=1}^{n} z'_{ij}}$$

We can now update with:

$$c_j = \sum_{i=1}^{n} \alpha_{ij} \cdot x^{(i)}$$

2) Let $<x^{(1)}, x^{(2)}>$ be the inner product of $x^{(1)}$ & $x^{(2)}$

$$\|x^{(1)} - x^{(2)}\|^2 = d(x^{(1)}, x^{(2)})^2 = <x^{(1)} - x^{(2)}, x^{(1)} - x^{(2)}> =$$

$$= <x^{(1)}, x^{(1)}> + <x^{(2)}, x^{(2)}> - 2<x^{(1)}, x^{(2)}>$$

3) $\|x^{(i)} - c_j\|^2 = <x^{(i)}, x^{(i)}> + <c_j, c_j> - 2<x^{(i)}, c_j>$

Initialize c_j = random$(x^{(i)})$

Iteration update c_j = mean$(x_j^{(i)})$

Support Vector Machines

(Func -> Fig)

1 -> 4

The regularization (inverse of C) is very high, hence this soft-margin linear SVM misclassifies several of its support vectors. The classifier is a straight line, since it’s just a linear kernel.

2 -> 3

The regularization (inverse of C) is very low, hence this soft-margin linear SVM correctly classifies all of its support vectors. The classifier is a straight line, since it’s just a linear kernel.

3 -> 5

Since k(x, z) is a second order function (squared), the resulting curve must be an ellipse or hyperbolic.

4 -> 6

The coefficient of -5 in the function $k(x, z) = \exp(-5||x - z||2)$, turns the exponent always into a negative number. e to the power of a large negative number gives a small number close to zero. The closer this number (the kernel value) is to zero the more squished together data points are in the kernel space. Hence there will be many samples that lie on the margin, i.e. there will be many support vectors for a hard margin kernel. Since there are more support vectors in figure 6 than in 1 and -5 results in a smaller number than – 1/5, 4 corresponds to 6 and 5 corresponds to 1.

5 -> 1

By the same argument that we have outline in 4 -> 6, 5 -> 1, as there are fewer support vectors in 1.