COMP0078PastExams

1 2022 Exam

1.1 1

Consider a dataset $(x_1, y_1), ..., (x_m, y_m) \in R \times R$, i.e., both x_i and y_i are real-valued scalars for all i. Suppose we wish to fit a linear+offset model,

$$\hat{y} = ax + b$$

(a)

Question Suppose we have a software library that performs least squares (without offset). Explain how we can use that software library to fit the linear+offset model.

Answer:

We can just subtract the offset b from y and proceed normal using least squares. Our regression coefficients become

$$(X^T X)^{-1} X^T (y - b)$$

where $X \in R^m$

(b) Question Suppose we use the least squares criterion to fit a linear model for the following dataset: $(x_1, y_1), ..., (x_m, y_m) \in R \times R$, by solving the following optimisation problem:

$$(a^*, b^*) = \text{argmin}_{a,b} \sum_{i=1}^{m} (y_i - ax_i + b)^2$$

Assume that the solution is unique. Which of the following statements are necessarily true? (If a statement is false you do not need to disprove it - but you need to prove true statements).

i) $\sum_{i=1}^{m} (y_i - a^* x_i + b^*) y_i = 0$
ii) $\sum_{i=1}^{m} (y_i - a^* x_i + b^*) x_i^2 = 0$
iii) $\sum_{i=1}^{m} (y_i - a^* x_i + b^*) x_i = 0$
iv) $\sum_{i=1}^{m} (y_i - a^* x_i + b^*)^2 = 0$

Answer:

We have to take the derivative of the loss function wrt a and b:

$$\frac{\partial}{\partial a}\sum_{i=1}^{m}(y_i - ax_i + b)^2 = 0 \tag{1}$$

$$\sum_{i=1}^{m}\frac{\partial}{\partial a}(y_i - ax_i + b)^2 = 0 \tag{2}$$

$$\sum_{i=1}^{m}2(y_i - ax_i + b)\frac{\partial}{\partial a}(y_i - ax_i + b) = 0 \tag{3}$$

$$\sum_{i=1}^{m}2(y_i - ax_i + b)(-x_i) = 0 \tag{4}$$

$$\sum_{i=1}^{m}-2(y_i - ax_i + b)x_i = 0 \tag{5}$$

$$\sum_{i=1}^{m}(y_i - ax_i + b)x_i = 0 \tag{6}$$

So this means that statement (iii) should be true.
And for b:

$$\frac{\partial}{\partial b}\sum_{i=1}^{m}(y_i - ax_i + b)^2 = 0 \tag{7}$$

$$\sum_{i=1}^{m}\frac{\partial}{\partial b}(y_i - ax_i + b)^2 = \sum_{i=1}^{m}2(y_i - a_i + b)\frac{\partial}{\partial b}(y - ax_i + b) = 0 \tag{8}$$

$$\sum_{i=1}^{m}2(y - ax_i + b)(+1) = 0 \tag{9}$$

$$\sum_{i=1}^{m}(y - ax_i + b) = 0 \tag{10}$$

Seems like only (iii) is true.

1.2 2

(a)
(i) **Question** State the ridge regression objective function.
Answer:
$\mathcal{E}_{\text{emp}_\lambda}(\mathbf{w} := \sum_{i=1}^{m}(y_i - \mathbf{w}^T\mathbf{x}_i)^2 + \lambda\sum_{l=1}^{n}w_l^2 = (\mathbf{y} - X\mathbf{w})^T(\mathbf{y} - X\mathbf{w}) + \lambda\|\mathbf{w}\|_2^2$

(ii) **Question** Give both the primal and dual solutions (no derivation required)
Answer:
Primal Solution: $\mathbf{w} = (X^TX + \lambda I_n)^{-1}X^Ty$
Dual solution: $\mathbf{w} = X(X^TX + \lambda I_m)^{-1}$
(iii) **Question** Discuss why depending on the setting, why one might select to use the primal solution or the dual solution.
Answer: Solving for $\mathbf{w}$ in the primal form requires $O(mn^2+n^3)$ operations while solving for α in the dual form requires $O(nm^2 + m^3)$. Hence, if $m << n$ it is more efficient to use the dual representation.
(b)

Suppose $K : X \times X \to R$ is a kernel function. Then consider the inequality

$$K(x_1, x_2)^2 \leq K(x_1, x_1) \times K(x_2, x_2)$$

(i-ii) **Question** Is this inequality true for all $x_1, x_2 \in X$?

Answer:

First let's define a kernel: **Kernel Definition** A function $k : \mathcal{X} \times \mathcal{X} \to R$ is a kernel if

1) k is symmetric: $k(x, z) = k(z, x)$

2) k gives rise to a positive semi-definite "Gram matrix", i.e., for any $m \in N$ and any $\mathbf{x} \in R^m$ chosen from $\mathcal{X}$, the Gram matrix $\mathbf{K}$ defined by $K_{il} = k(x_i, x_l)$ is positive semidefinite.

Another way to show that a matrix K is positive semi-definite is to show that

$$\forall \mathbf{c} \in R^m, \mathbf{c}^T \mathbf{K} \mathbf{c} \geq 0$$

Now, we want to show that $K(x_1, x_2)^2 \leq K(x_1, x_1) \times K(x_2, x_2)$

We have that the Gram matrix is

$$\mathbf{K} = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) \\ k(x_2, x_1) & k(x_2, x_2) \end{pmatrix}$$

is positive semi-definite whenever $\mathbf{c}^T \mathbf{K} \mathbf{c} \geq 0 \forall c$. Choose in particular $\mathbf{c} = \begin{pmatrix} k(x_2, x_2) \\ -k(x_1, x_2) \end{pmatrix}$

Then since $\mathbf{K}$ is positive semi-definite,

$$0 \leq \mathbf{c}^T \mathbf{K} \mathbf{c} = \big(k(x_2, x_2) k(x_1, x_1) - k(x_1, x_2)^2\big) k(x_2, x_2)$$

So we must then have that $k(x_1, x_2)^2 \leq k(x_1, x_1) k(x_2, x_2)$

2 3

Given the data set,

$((1, 1), +1), ((2, 2), +1), ((2, 0), +1), ((0, 0), -1), ((1, 0), ((0, 1), -1)$

(a) **Question** Plot the dataset as well as the maximal margin hyperplane. Give the equation of the maximum margin hyperplane. (Note: This may be done by inspection rather than derivation and proof). Finally, identify the support vector(s).

Answer:

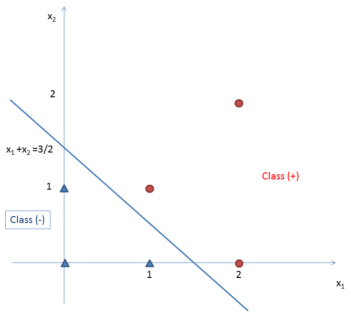

Figure 1

By inspection it seems like our line passes through $(0, 1.5)$ and $(1.5, 0)$. So we find the slope:

$$m = \frac{0 - 1.5}{1.5 - 0} = \frac{-1.5}{1.5} = -1$$

So we have that

$$y - y_1 = m(x - x_1) \tag{11}$$
$$y - 1.5 = -1(x - 0) \tag{12}$$
$$y - 1.5 = -x \tag{13}$$
$$y + x = 1.5 \tag{14}$$

So the maximum margin hyperplane is $x_1 + x_2 = 1.5$
The support vectors are:
For class +: $(1, 1), (2, 0)$
For class -: $(1, 0), (0, 1)$
(b) **Question** Suppose we remove a datapoint from the above training set. How will the maximum margin change?
(i) Strictly increases
(ii) Strictly decreases
(iii) Stays the same
(iv) Depends on the removed data point. If you choose this then explain how it depends on the data point.
Answer:
It depends on the datapoint. If you remove either $(1, 0)$ or $(1, 1)$ the margin will increase, however if you remove either $(1, 1)$ or $(2, 0)$ then it will stay the same. And for any other point which isn't a support vector - it will stay the same, since they don't affect the margin.

3 4

(a) Question Given the data set with five examples,

$$((1, 1), +1), ((1, -1), +1), ((-1, 1), +1), ((-1, -1), -1), ((0, 0), -1)$$

Plot the dataset. Consider training a classifier with Adaboost using decision stumps. Indicate which example(s) have their weights increased after a single iteration of boosting. Explain why.
Answer:

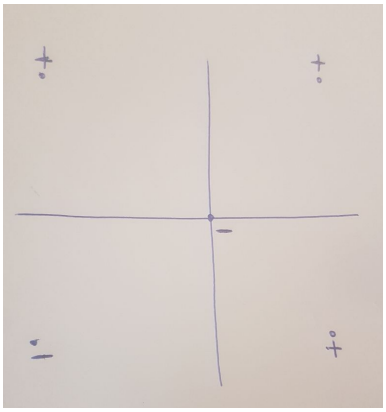

Figure 2

There are 2 decision stumps which can achieve the minimal error at the first iteration. The first one splits the data with $x_1 \leq 0$ and the second one splits with $x_2 \leq 0$. The algorithm chooses between these 2 stumps uniformly.

In case the first stumps is chosen, we split with $x_1 \leq 0$ and the samples which are less than or equal to 0 will be classified as negative, and the sample larger than 0 will be classified as positive. Since the point $((-1, 1), +1)$ is classified as negative although it's positive, it will its weight increased.

In case the second stump is chosen, we split with $x_2 \leq 0$ and the samples which are less than or equal to 0 will be classified as negative, and the samples larger than 0 will be classified as positive. Since the point $((1, -1), +1)$ is classified as negative although it's positive, it will its weight increased.

(b) Question

Suppose AdaBoost is run on m training examples, and suppose on every round that the weighted training error ϵ_t of the t-th weak hypothesis is at most $\frac{1}{2} - \gamma$, for some $\gamma > 0$. What is an upper bound on the number of iterations, T (if such a number exists), until the combined hypothesis H is always consistent with the m training examples, i.e., achieves zero (unweighted) training error? You answer should only be expressed in terms of m and γ.

Answer:

When we have 1 misclassified example, the training error is $\frac{1}{m}$. So in order to get zero training error, we need the training error to be less than $\frac{1}{m}$. We know that from lecture notes

$$H \leq \exp(-2T\gamma^2)$$

so we want $\exp(-2T\gamma^2) < \frac{1}{m}$:

$$\exp(-2T\gamma^2) < \frac{1}{m} \tag{15}$$

$$-2T\gamma^2 < \ln(\frac{1}{m}) \tag{16}$$

$$T > -\frac{\ln(\frac{1}{m})}{2\gamma^2} \tag{17}$$

$$T > \frac{-(0 - \ln(m))}{2\gamma^2} \tag{18}$$

$$T > \frac{\ln(m)}{2\gamma^2} \tag{19}$$

4 6

Consider the following questions in the context of the model of PAC learning.

(i) Question Give an example of an infinite hypothesis class with **finite** sample complexity.

Answer:

Just like in the lecture notes, define $\text{sign}(z) = +1$ if $z \geq 0$ and -1 otherwise. Let $X = R$ and $\mathcal{H} = \{h_\theta(x) = \text{sign}(x - \theta) : \theta \in R\}$. It's clear that e.g. $\{0\}$ is shattered, since for $\theta = 1$, $h_\theta(0) = -1$ and for $\theta = -1$ we get $h_\theta(0) = +1$, so we represent all of the possible outcomes the point 0 can be classified to. However for more than 1 point it's not possible as we can have $(-, -)$, $(-, +)$, $(+, +)$ but we can't have $(+, -)$. So $VC(\mathcal{H}) = 1$.

So, by the Fundamental Theorem of Statistical Learning, the sample complexity is finite and the hypothesis space is infinite.

(ii) Question Give an example of an infinite hypothesis class with **infinite** sample complexity.

Answer:

Let $\mathcal{X} = R$ and $\mathcal{H} = R$. For $w \in R$ a hypothesis is given by

$$h_w(x) = \text{sign}(\sin(wx))$$

(iii) Question For a finite hypothesis class what does the cardinality of the hypothesis class reveal about the sample complexity. [Recommended answer length 2-4 sentences]

Answer:

Let $m_{\mathcal{H}}$ be the sample complexity of learning $\mathcal{H}$. Then for $(\epsilon, \delta) \in (0,1)^2$,

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}$$

(iv) Question For the infinite hypothesis class of "part ii" sketch an argument and provide intuitions on why the sample complexity is infinite. [Recommended answer length 4-8 sentences]

Answer:

For all m, the set $S = \{2^1, 2^2, ..., 2^m\}$ is shattered by h. To see this, let $w = -\pi * (0.y_1 y_2 ... y_m)$ be a decimal binary encoding of a set of desired labels, converting -1 to 0. Essentially each x_i bit shifts w to produce the desired label as a result of the fact that $\text{sign}(\sin(\pi z)) = (-1)^{\lfloor z \rfloor}$. This the VC dimension of this hypothesis class is infinite.