Completed

Principles of Machine Learning

Exercises

Victor Verreet *victor.verreet@cs.kuleuven.be*
Laurens Devos *laurens.devos@cs.kuleuven.be*

Fall, 2021

Exercise Session 4: Learning Theory and Support Vector Machines

4.1 The RPNI Algorithm

Let $\{aa, ab, ba\}$ be the positive examples and $\{baa\}$ the negative examples of strings over the alphabet $\Sigma = \{a, b\}$. Apply the RPNI algorithm to find a DFA that correctly labels these instances. Which regular expression does this correspond to?

4.2 VC-Dimension

Consider the instance space $\mathbb{R}^2$ and take as a hypothesis space the set of all hypotheses of the form "everything inside rectangle R is positive and everything outside it is negative", where R can be any axis-parallel rectangle. This means that the sides of the rectangle are parallel to the X- and Y-axis. Show that the VC-dimension of this hypothesis space is at least 3. Optionally show that the VC-dimension of this hypothesis space is at least 4.

Now consider a similar hypothesis space, where for each hypothesis that states "everything inside rectangle R is positive and everything outside it is negative" there is also a hypothesis "everything outside R is positive and everything inside is negative". Show that the VC-dimension of this hypothesis space is at least 4.

4.3 The Pizza Problem

Suppose pizzas can be composed of the following 8 ingredients: *cheese, ham, tomatoes, mushrooms, peppers, salami, olives* and *shrimp*. Further assume that the concept "pizzas that Wayne likes" can be expressed as a conjunction of "contains I" literals, with I a single ingredient. For instance, "Wayne likes all pizzas that contain peppers and salami" is a possible concept, whereas "Wayne likes all pizzas that contain peppers and no olives" is not. Show that the VC-dimension of this hypothesis space is at least 3 and at most 8.

4.4 Sample Complexity

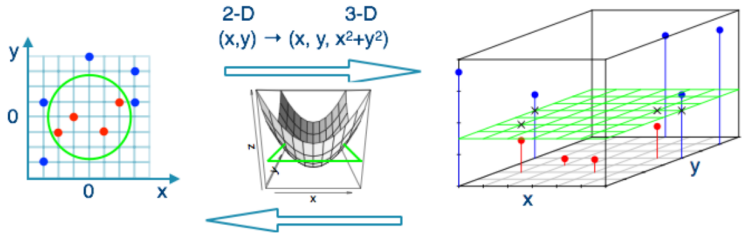
Using the concept of VC-dimension, compute a lower bound for the number of examples that may be needed to train a 2-input perceptron such that with 90% certainty it learns a hypothesis with true error smaller than 5%.

Hint: The first step is to find a lower bound for the VC-dimension of a perceptron. Then, look in the slides for a formula linking VC-dimension to number of required samples.

4.5 SVM Kernel Trick: Mapping to Higher Dimensions

To learn a non-linear SVM, the input data is transformed to a higher-dimensional space and a linear SVM is learned in this new space. The linear separator in the higher-dimensional space then corresponds to a non-linear separator in the original space. The hope is that the data is actually linearly separable after the transformation.

The following visual example was given in the slides.



We also saw that in the dual formulation, the transformation only appeared in dot-product pairs: $\Phi(x) \cdot \Phi(x')$ for two input instances x and x'. As a result, we saw that it is not necessary to explicitly specify Φ, using just a kernel function $K(x, x')$ is sufficient. One requirement for such a kernel is that it corresponds to a dot product in a transformed space.

Consider the following kernel K:

$$K(v_1, v_2) = (v_1 \cdot v_2 + 1)^2,$$

where v_1 and v_2 are two vectors in $\mathbb{R}^2$. Show that this kernel corresponds to

$$\langle \Phi(v_1), \Phi(v_2)) \rangle_F$$

where Φ is the following transformation:

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^{3\times 3} : v = (x, y) \mapsto \Phi(v) = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \begin{bmatrix} x & y & 1 \end{bmatrix},$$

and $\langle A, B \rangle_F$ is the Frobenius inner product defined on two matrices A, and B:

$$\langle A, B \rangle_F = \text{trace}(A^T B) = \sum_i \sum_j A_{ji} B_{ji}.$$

Note that Φ maps a vector v, after extending it with the constant 1, to its outer product.

4.6 Feature Selection

Assume we want to classify instances into three classes c_1, c_2 and c_3. Our original input has three features a_1, a_2 and a_3 of which we want to keep only two. Feature a_1 is similar for instances of classes c_1 and c_2, but different for c_3 instances. Feature a_2 is similar for instances of classes c_2 and c_3, but different for c_1 instances. Instances of all three classes have similar a_3. Let f_i be the fraction of instances of class c_i in the training dataset. We know that $f_1 > f_2 > f_3$.

When performing greedy forward feature selection, which feature will get selected first? Which second? Why?

4.7 Fatigue Prediction

Take a look at this paper[1] which can be found at `https://dl.acm.org/citation.cfm?id=3219864`. The aim of this research is to predict using machine learning techniques how fatigued runners are after performing with the aim of preventing injuries.

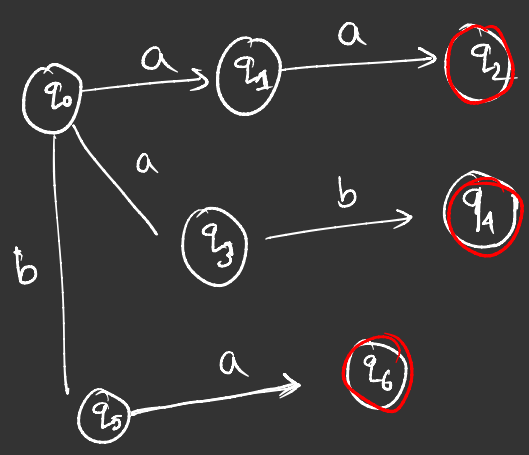
[1] Fatigue Prediction in Outdoor Runners Via Machine Learning and Sensor Fusion; Tim Op De Beéck, Wannes Meert, Kurt Schütte, Benedicte Vanwanseele, Jesse Davis; 2018

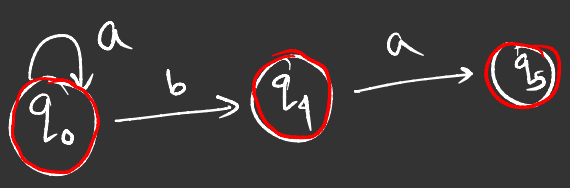
For this purpose, five accelerometers were attached to the runners, one at the back, one at each wrist and one at each leg. The data of all these sensors can be combined to predict the runner's fatigue. It is also possible to fuse this data together before using it to predict. Furthermore some general statistical features and some more advanced sport science features are calculated on these temporal data.

Read the paper and try to think about the following questions:

1. Which sensor locations are most useful?
2. Does fusing data affect prediction and run time performance?
3. Are the advanced sport science features useful?

4.1

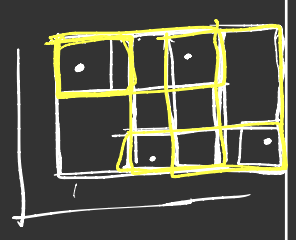




4.2 let $X = \{(x_1,y_1), (x_2,y_2), (x_3,y_3)\}$ be three pts. in R^2 s.t $x_1 < x_2 < x_3$, $y_3 < y_2, y_1$

Case I

When 2 positive & 1 negative



4.3

$x = (0,0,0,0,1,1,1,1)$ 0

$y = (0,0,1,1,0,0,1,1)$ 1

$z = (0,1,0,1,0,1,0,1)$ 1

$4\alpha(x) + 2\alpha(y) + \alpha(z)$

4.4 $\epsilon = 0.05$

$1-\delta = 0.9 \Rightarrow \delta = 0.1$

$|H| = 3$

$\frac{13 \times 100}{5}^{20}$

$|T| \geq \frac{1}{0.05}\left(4\log_2\frac{2}{0.1} + 8\times 3\log_2\frac{13}{0.05}\right)$

$\geq \frac{1}{0.05}\left(4\log_2 20 + 24 \times \log 260\right)$

≥ 4197

4.5

$v_1 = (x_1, y_1)$

$v_2 = (x_2, y_2)$

$K(v_1, v_2) = (x_1x_2 + y_1y_2 + 1)^2$

$= x_1^2x_2^2 + y_1^2y_2^2 + 1 + 2x_1x_2y_1y_2 + 2x_1x_2 + 2y_1y_2$

$\phi(v_1)^T \phi(v_2)$

$= \begin{pmatrix} x_1 \\ y_1 \\ 1 \end{pmatrix}(x_1\ y_1\ 1)\begin{pmatrix} x_2 \\ y_2 \\ 1 \end{pmatrix}(x_2\ y_2\ 1)$

$= \begin{pmatrix} x_1 \\ y_1 \\ 1 \end{pmatrix}(x_1x_2 + y_1y_2 + 1)(x_2\ y_2\ 1)$

$= \begin{pmatrix} x_1^2x_2 + x_1y_1y_2 + x_1 \\ x_1y_1x_2 + y_1^2y_2 + y_1 \\ x_1x_2 + y_1y_2 + 1 \end{pmatrix}(x_2\ y_2\ 1)$

$= \begin{pmatrix} x_1^2x_2^2 + x_1x_2y_1y_2 + x_1x_2 & \cdots \\ \cdots & x_1y_1x_2y_2 + y_1^2y_2^2 + y_1y_2 \\ \cdots & \cdots \end{pmatrix}$

$\cdots \quad x_1x_2 + y_1y_2 + 1\Big)$

$y_1^2y_2^2 + x_1^2x_2^2 + 2x_1x_2y_1y_2 + 2x_1x_2 + 2y_1y_2 + 1$

4.6

①	a_1	$f_1 + f_3$	$q_2 = f_1 + f_2$
②	a_2	$f_1 + f_2 + f_3$	q_1