MTAT.03.227 Machine Learning
Practice session 5
Distance-Based and Kernel Methods

October 07-11, 2019

Exercise 0. Practice Slides

Go through the practice material provided here and solve the following tasks.

Exercise 1. Primal Vs Dual Form of Perceptron

Algorithm Perceptron(D, η) – train a perceptron for linear classification.

Input : labelled training data D in homogeneous coordinates; learning rate η.
Output : weight vector $\mathbf{w}$ defining classifier $\hat{y} = \text{sign}(\mathbf{w} \cdot \mathbf{x})$.

$\mathbf{w} \leftarrow \mathbf{0}$; // Other initialisations of the weight vector are possible
converged$\leftarrow$false;
while *converged* = false **do**
    converged$\leftarrow$true;
    for $i = 1$ to $|D|$ **do**
        if $y_i \mathbf{w} \cdot \mathbf{x}_i \leq 0$
        then
            $\mathbf{w} \leftarrow \mathbf{w} + \eta y_i \mathbf{x}_i$;
            converged$\leftarrow$false;
        end
    end
end

Algorithm DualPerceptron(D) – perceptron training in dual form.

Input : labelled training data D in homogeneous coordinates.
Output : coefficients α_i defining weight vector $\mathbf{w} = \sum_{i=1}^{|D|} \alpha_i y_i \mathbf{x}_i$.

$\alpha_i \leftarrow 0$ for $1 \leq i \leq |D|$;
converged$\leftarrow$false;
while *converged* = false **do**
    converged$\leftarrow$true;
    for $i = 1$ to $|D|$ **do**
        if $y_i \sum_{j=1}^{|D|} \alpha_j y_j \mathbf{x}_i \cdot \mathbf{x}_j \leq 0$ **then**
            $\alpha_i \leftarrow \alpha_i + 1$;
            converged$\leftarrow$false;
        end
    end
end

Given the pseduocode for the perceptron algorithm above in both primal form and dual form.

(a) What is the formula for making classification predictions with the primal form perceptron ($\hat{y} =$?) ?

(b) How is the dual form of the perceptron algorithm different from the primal form?

(c) The final weights of linear classifier are $\mathbf{W} = \sum_{i=1}^{n} \alpha_i \eta y_i \mathbf{X_i}$. Explain the meaning of each quantity in this formula.

(d) How to write out the formula for making classification predictions with the dual form perceptron ($\hat{y} =$?) ? Note that it should not contain $\mathbf{W}$, as the dual form does not explicitly calculate $\mathbf{W}$.

Exercise 2. Kernel Trick

Suppose that we have a linearly non-separable dataset with two features. So, we are going to try out one of the feature transformation functions ($\mathbf{X} \rightarrow \phi(\mathbf{X})$) and see how it works (Practice Session Slides 16-18).

(a) What are the new features introduced by the feature transformation function?

(b) Is it possible to apply the kernel trick instead of feature transformation in the perceptron algorithm (Primal Form / Dual Form)?

(c) Let us suppose that we have a dataset where, unfortunately, the new features introduced by the feature transformation function / kernel used in (a) are still not separating our classes. So, we are going to try out another 2nd degree polynomial kernel $\mathbf{K}(\mathbf{X}, \mathbf{X}') = (\mathbf{X} \cdot \mathbf{X}' + r)^d$ where $r = 2,\ d = 2$.

Find the transformation function $\phi(\mathbf{X})$ such that $\mathbf{K}(\mathbf{X_1}, \mathbf{X_2}) = \phi(\mathbf{X_1}) \cdot \phi(\mathbf{X_2})$

(d) What are the new features added by the new kernel?

(e) Consider that we have two instances of two feature vectors $\mathbf{X_1} = (6,3)$, and $\mathbf{X_2} = (5,4)$. Calculate the result of $\phi(\mathbf{X_1}) \cdot \phi(\mathbf{X_2})$ and measure the number of operations needed to achieve that without the kernel trick.

(f) Use the above polynomial kernel $\mathbf{K}(\mathbf{X},\mathbf{X}') = (\mathbf{X} \cdot \mathbf{X}')^2$ to calculate the value of $\phi(\mathbf{X_1}) \cdot \phi(\mathbf{X_2}) = \mathbf{K}(\mathbf{X_1},\mathbf{X_2})$, and measure the number of operations needed to achieve that. Compare the number of operations with the previous subtask.

Exercise 3. Degree-3 Kernel (Homework)

For data with 2 features, find the transformation function $\phi(\mathbf{X})$ corresponding to a polynomial kernel of degree 3 and no intercept, that is $\mathbf{K}(\mathbf{X},\mathbf{X}') = (\mathbf{X} \cdot \mathbf{X}' + r)^d$ where $r = 0,\ d = 3$ such that $\mathbf{K}(\mathbf{X},\mathbf{X}') = \phi(\mathbf{X}) \cdot \phi(\mathbf{X}')$

Exercise 4. Gaussian (RBF) Kernel

Given the formula for Gaussian kernel as $\mathbf{K}(\mathbf{X},\mathbf{X}') = \exp\left(\frac{-\|\mathbf{X}-\mathbf{X}'\|^2}{2\alpha}\right)$

(a) What is the kernel value when $\mathbf{X}$ and $\mathbf{X}'$ are very similar / very different ?

(b) What is the dimension of the feature space constructed by this kernel ?