Introduction to Machine Learning Course

Short HW3 – SVM, Optimization, and PAC learning

Submitted individually by Wednesday, 31.07, at 23:59.

You may answer in Hebrew or English and write on a computer or by hand (but be clear).

Please submit a PDF file named like your ID number, e.g., 123456789.pdf.

Bonus (maximal grade is 100): Writing on a computer (using LyX/LaTeX, Word + Equation tool, etc.) = 2 pts.

חיילים בשירות מילואים ממושך המעוניינים לקבל פטור מאחת השאלות במטלה (לבחירתנו) מוזמנים לפנות למייל הקורס.

1. Define $\mathcal{H} = \{x \mapsto sign(w^t x) : w \in \mathbb{R}^d\}$, the hypothesis class of homogeneous linear classifiers.

1.1. In Tutorial 05, we said that the VC-dimension of homogeneous linear classifiers is $\geq d$. Provide a rigorous proof for this statement.

1.2. Prove that $VCdim(\mathcal{H})$ is exactly d by proving that $VCdim(\mathcal{H}) < d + 1$.

Hint: Any set of $\{x_1, \ldots, x_{d+1}\}$ vectors in $\mathbb{R}^d$ is linearly dependent, and at least one vector in the set (w.l.o.g x_{d+1}) satisfies $x_{d+1} = \sum_{i=1}^{d} z_i x_i$ for some scalars $z_1, \ldots, z_d \in \mathbb{R}$ with at least some scalar that is not equal to 0.

2. Let $\phi: \mathcal{X} \to \mathbb{R}^{n_1}, \phi': \mathcal{X} \to \mathbb{R}^{n_2}$ be two feature mappings where $n_1, n_2 \in \mathbb{N}$.

Let $K, K': (\mathcal{X} \times \mathcal{X}) \to \mathbb{R}$ be two valid kernels defined as:

$$K(u,v) = \langle \phi(u), \phi(v) \rangle = \sum_{i=1}^{n_1} \phi_i(u)\phi_i(v)\,, \quad K'(u,v) = \langle \phi'(u), \phi'(v) \rangle = \sum_{j=1}^{n_2} \phi'_j(u)\phi'_j(v).$$

Prove that $G(u,v) \triangleq K(u,v) \cdot K'(u,v)$ is a valid kernel. That is, propose a feature mapping $\psi: \mathcal{X} \to \mathbb{R}^{n_3}$ for some $n_3 \in \mathbb{N}$, such that $G(u,v) = \langle \psi(u), \psi(v) \rangle$.

Hint: You should use $n_3 = n_1 \cdot n_2$.

3. For a given parameter $\gamma > 0$, define the Gaussian Kernel for 1-D input in the following manner:

$$K: \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \qquad K(a,b) = \exp(-\gamma(a-b)^2)$$

3.1. Provide a feature mapping $\phi: \mathbb{R} \to \mathbb{R}^p$ with $p \in \mathbb{N} \cup \{\infty\}$, and prove that K is indeed a valid kernel.

Hint: $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$

3.2. Assume that you are given a very large dataset with 1-D samples. We would like to apply the Gaussian Kernel to train a classifier on the dataset. Would it be better to optimize the **primal problem** with the feature mapping you found, or is it better to optimize the **dual problem** with the kernel that we defined? Is it even possible? Explain.

4. **Refute** (with a simple example)**:** Let $f, g: \mathbb{R} \to \mathbb{R}$ be two convex functions.

The composition $h \triangleq f \circ g$ (that is, $h(x) = f\big(g(x)\big)$) is also a convex function.

5. We will now prove that the following Soft-SVM problem is convex:

$$\boxed{\underset{w\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m} \max\{0,1-y_i\cdot w^\top x_i\}+\lambda\|w\|_2^2}$$

Let $f, g: C \to \mathbb{R}$ be two convex functions defined over a convex set C.

Lemma (no need to prove)**:** $q(z) \triangleq \max\{f(z),\ g(z)\}$ is convex w.r.t z.

Lemma (no need to prove)**:** the sum of any number of convex functions is convex.

5.1. Prove (by definition): Given a constant $\alpha \in \mathbb{R}_{\geq 0}$, the function $\alpha f(z)$ is convex w.r.t z.

5.2. Using a rule from Tutorial 07, conclude that $\max\{0, 1-y_i\boldsymbol{w}^\top\boldsymbol{x}_i\}$ is convex w.r.t $\boldsymbol{w}$.

5.3. Using the above (and properties from Tutorial 07), conclude that the Soft-SVM optimization problem is convex w.r.t $\boldsymbol{w}$.