STATISTICAL MACHINE LEARNING (WS2017)
SEMINAR 3

Assignment 1. [1] There have been 58 US presidential elections. Let us see each county's voting outcome as a predictor $h\colon \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} = \{1, 2, \ldots\}$ is the set of election indices and $\mathcal{Y}$ is a set of all presidential candidates. Assume that the sequence of elected presidents { "George Washington", "George Washington", ... , "Barack Obama", "Donald Trump" } is a realization of i.i.d. random variables with unknown distribution $p(y)$. There are $|\mathcal{H}| = 3100$ US counties. Suppose there is a county h' which has always correctly predicted the elected US president. What is the probability that this county will not predict the correct president in the future elections with confidence at least 95%?

Assignment 2. [2] Let us consider the space of all linear classifiers mapping $\boldsymbol{x} \in \mathbb{R}^d$ to $\{-1, +1\}$, that is

$$\mathcal{H} = \left\{ h(\boldsymbol{x}; \boldsymbol{w}, b) = \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b) \mid (\boldsymbol{w}, b) \in (\mathbb{R}^d \times \mathbb{R}) \right\} .$$

Show that the VC dimension of $\mathcal{H}$ is $d + 1$.

Assignment 3. Consider a hypothesis space of classifiers

$$\mathcal{H} = \left\{ h(x; a) = \operatorname{sign}(\sin(ax)) \mid a \in \mathbb{R} \right\} .$$

That is, each $h \in \mathcal{H}$ is determined by a single parameter $a \in \mathbb{R}$ and it maps real valued input $x \in \mathbb{R}$ to a set of hidden labels $\{+1, -1\}$ based on the sign of the score $\sin(xa)$. Show that the VC dimension of $\mathcal{H}$ is infinite.

Hint: Show that for arbitrary set of labels $\{y^i \in \{+1, -1\} \mid i = 1, \ldots, m\}$ the inputs $\{x^i = 10^{-i} \mid i = 1, \ldots, m\}$ can be predicted correctly by $h(x; a)$ with

$$a = \pi \left(1 + \frac{1}{2} \sum_{i=1}^{m} (1 - y^i) 10^i \right)$$

Assignment 4. Assume we are given a training set of examples $\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \{+1, -1\}) \mid i = 1, \ldots, m\}$ which is known to be linearly separable with respect to a feature map $\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^n$. In this case, we can find parameters $(\boldsymbol{w}, b) \in \mathbb{R}^{n+1}$ of a linear classifier $h(x; \boldsymbol{w}, b) = \operatorname{sign}(\langle \boldsymbol{\phi}(x), \boldsymbol{w} \rangle + b)$ which has zero training error by the Perceptron algorithm:

(1) $\boldsymbol{w} \leftarrow 0, b \leftarrow 0$

(2) Find an example $(x^u, y^u) \in \mathcal{T}^m$ whose label is incorrectly predicted by the current classifier, that is $h(x^u; \boldsymbol{w}, b) \neq y^u$.

[1] Adopted from Xiaojin Zhu http://pages.cs.wisc.edu/~jerryzhu/teaching.html

[2] This assignment is relatively complicated. You may skip it you find it too difficult.

(3) If all examples are classified correctly exit the algorithm. Otherwise update the parameters by

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + y^u \boldsymbol{\phi}(x^u) \qquad \text{and} \qquad b \leftarrow b + y^u$$

and go to Step 2.

Assume that you cannot evaluate the feature map $\boldsymbol{\phi}(x)$ because it is either unknown or its evaluation is expensive. However, you know how to cheaply evaluate a kernel function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k(x, x') = \langle \boldsymbol{\phi}(x), \boldsymbol{\phi}(x') \rangle$, $\forall x, x' \in \mathcal{X}$. Show that you can still use the Perceptron algorithm to find a linear classifier with zero training error and that you can evaluate this classifier on any $x \in \mathcal{X}$.

Assignment 5. Let the input observation be a vector $\boldsymbol{x} \in \mathbb{R}^d$. Let us consider a feature map $\boldsymbol{\phi}_q\colon \mathbb{R}^d \to \mathbb{R}^n$, $n = d^q$, whose entries are all possible q-th degree ordered products of the entries of $\boldsymbol{x}$. For example, if $\boldsymbol{x} = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ and $q = 2$ then

$$\boldsymbol{\phi}_q(\boldsymbol{x}) = \begin{pmatrix} x_1x_1 \\ x_2x_1 \\ x_3x_1 \\ x_1x_2 \\ x_2x_2 \\ x_3x_2 \\ x_1x_3 \\ x_2x_3 \\ x_3x_3 \end{pmatrix}$$

a) Show that for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ we can compute the dot product between $\boldsymbol{\phi}_q(\boldsymbol{x})$ and $\boldsymbol{\phi}_q(\boldsymbol{x}')$ as

$$\langle \boldsymbol{\phi}_q(\boldsymbol{x}), \boldsymbol{\phi}_q(\boldsymbol{x}') \rangle = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle^q \ ,$$

that is, as the dot product of the original vectors $\boldsymbol{x}$ and $\boldsymbol{x}'$ powered to q.

b) Consider a slightly different feature map $\boldsymbol{\phi}'\colon \mathbb{R}^d \to \mathbb{R}^{d(d+1)/2}$ whose entries are

$$\begin{array}{rcllllll} \boldsymbol{\phi}'(\boldsymbol{x}) & = & (& x_1^2, & \sqrt{2}x_1x_2, & \sqrt{2}x_1x_3, & \ldots, & \sqrt{2}x_1x_d, \\ & & & & x_2^2, & \sqrt{2}x_2x_3, & \ldots, & \sqrt{2}x_2x_d, \\ & & & & & & & \vdots \\ & & & & & & & x_d^2 \qquad)^T \ , \end{array}$$

so that the features correspond to all possible products of unordered pairs of entries from $\boldsymbol{x}$, and the products of different entries are multiplied by a constant factor $\sqrt{2}$. For example, if $\boldsymbol{x} = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ then

$$\boldsymbol{\phi}'(x) = (x_1^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, x_2^2, \sqrt{2}x_2x_3, x_3^2)^T \ .$$

This feature map defines a kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{\phi}'(\boldsymbol{x}), \boldsymbol{\phi}'(\boldsymbol{x}') \rangle$ referred to as the homogeneous polynomial kernel of degree 2. Show that the kernel value equals to the square of the dot product of the input vectors, that is prove the identity

$$k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{\phi}'(\boldsymbol{x}), \boldsymbol{\phi}'(\boldsymbol{x}') \rangle = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle^2 \,, \qquad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d \,.$$

Hint: Exploit the relation between $\boldsymbol{\phi}(\boldsymbol{x})$ *and* $\boldsymbol{\phi}'(\boldsymbol{x})$.