Exercises
Advanced Machine Learning
Fall 2019

Series 5, Nov 11th, 2019 (Support Vector Machines)

Machine Learning Laboratory
Dept. of Computer Science, ETH Zürich
Prof. Joachim M. Buhmann
Web https://ml2.inf.ethz.ch/courses/aml/

Teaching assistant: **Piazza (Francesco Locatello)**
locatelf@ethz.ch

Solution 1 (Warm-up: Kernel Function):

$$
\begin{aligned}
k_m(\vec{x},\vec{y}) &= \sum_{h,h'} k\big((\vec{x},h),(\vec{y},h')\big)\, p(h \mid \vec{x})\, p(h' \mid \vec{y}) \\
&= \sum_{h,h'} k\big((\vec{x},h),(\vec{y},h')\big)\, k_1\big((h,\vec{x}),(h',\vec{y})\big) && (k_1 \text{ is a kernel on } (\vec{x},h) \text{ pairs}) \\
&= \sum_{h,h'} k_2\big((h,\vec{x}),(h',\vec{y})\big) && (\text{The product of two kernels is also a kernel}) \\
&= \sum_{h,h'} \Phi_1(h,\vec{x})^\top \Phi_1(h',\vec{y}) \\
&= \left[\sum_h \Phi_1(h,\vec{x})\right]^\top \left[\sum_{h'} \Phi_1(h',\vec{y})\right] \\
&= \Phi_2(\vec{x})^\top \Phi_2(\vec{y}) \\
&= k_3(\vec{x},\vec{y}).
\end{aligned}
$$

Solution 2 (SVMs as Nearest Neighbor Classifiers):

Consider some $\vec{x} \in \mathbb{R}^d$ and let $\vec{x}_p$ denote its *unique* nearest neighbor amongst $\{\vec{x}_1,\ldots,\vec{x}_n\}$. Furthermore, let $\vec{x}_q \in \{\vec{x}_1,\ldots,\vec{x}_n\}$ denote the "second nearest neighbor" to $\vec{x}$. Of course, $\vec{x}_q$ may not be unique – in that case, we choose arbitrarily any of the candidate points as $\vec{x}_q$. Since by assumption $\alpha_i = 1$ for all $i = 1,\ldots,n$, we have that the SVM prediction is given by:

$$
\begin{aligned}
f(\vec{x}) &= \mathsf{sign}\left(\sum_{i=1}^n y_i \exp\left(-\frac{\|\vec{x}-\vec{x}_i\|^2}{h^2}\right)\right) \\
&= \mathsf{sign}\left(y_p \exp\left(-\frac{\|\vec{x}-\vec{x}_p\|^2}{h^2}\right) + \sum_{j=1, j\neq p}^n y_j \exp\left(-\frac{\|\vec{x}-\vec{x}_j\|^2}{h^2}\right)\right). \qquad (1)
\end{aligned}
$$

Observe that if we can find conditions on h which guarantee that the following inequality holds,

$$
\left| y_p \exp\left(-\frac{\|\vec{x}-\vec{x}_p\|^2}{h^2}\right)\right| > \left|\sum_{j=1, j\neq p}^n y_j \exp\left(-\frac{\|\vec{x}-\vec{x}_j\|^2}{h^2}\right)\right|, \qquad (2)
$$

then we have that $f(\vec{x}) = \mathsf{sign}(y_p)$ and hence the predicted label will be the same as that of a 1-nearest neighbor (NN) classifier. We will now work backwards, searching for conditions that make 2 hold. We start from the following relations involving its left- and right-hand terms:

$$\left| \sum_{j=1, j\neq p}^{n} y_j \exp\left(-\frac{\|\vec{x}-\vec{x}_j\|^2}{h^2}\right) \right| \leq (n-1)\exp\left(-\frac{\|\vec{x}-\vec{x}_q\|^2}{h^2}\right) \tag{3}$$

$$\left| y_p \exp\left(-\frac{\|\vec{x}-\vec{x}_p\|^2}{h^2}\right) \right| = \exp\left(-\frac{\|\vec{x}-\vec{x}_p\|^2}{h^2}\right). \tag{4}$$

Therefore, a sufficient condition for 2 is

$$\begin{aligned} & \exp\left(-\frac{\|\vec{x}-\vec{x}_p\|^2}{h^2}\right) > (n-1)\exp\left(-\frac{\|\vec{x}-\vec{x}_q\|^2}{h^2}\right) \\ \iff & \exp\left(\frac{\|\vec{x}-\vec{x}_q\|^2 - \|\vec{x}-\vec{x}_p\|^2}{h^2}\right) > (n-1) \\ \iff & \sqrt{\frac{\|\vec{x}-\vec{x}_q\|^2 - \|\vec{x}-\vec{x}_p\|^2}{\log(n-1)}}) =: h_0 > h \end{aligned} \tag{5}$$

Hence, for all $h < h_0$ we have that $f(\vec{x}) = \text{sign}(y_p) =$ the label of nearest neighbor of $\vec{x}$.

Solution 3 (Dual Formulation for Structural SVM):

Let $\mathbb{K}_i = \mathbb{K} \setminus \{z_i\}$. The Lagrangian is

$$\mathcal{L}(\vec{w}, \xi, \alpha, \beta) = \frac{1}{2}\vec{w}^\top \mathbf{w} + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \sum_{z_j \in \mathbb{K}_i} \alpha_{ij}(\mathbf{w}^\top \Psi_i(z_j) - \Delta_i(z_j) + \xi_i) - \sum_{i=1}^{n} \beta_i \xi_i. \tag{6}$$

The stationary conditions are

$$\nabla_{\vec{w}} \mathcal{L} \overset{!}{=} 0 \implies \vec{w} = \sum_{i=1}^{n} \sum_{z_j \in \mathbb{K}_i} \alpha_{ij} \Psi_i(z_j) \tag{7}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} \overset{!}{=} 0 \implies C = \beta_i + \sum_{z_j \in \mathbb{K}_i} \alpha_{ij} \quad i = 1, 2, \ldots, n. \tag{8}$$

Note that the second one together with $\beta_i \geq 0$ implies $C \geq \sum_{z_j \in \mathbb{K}_i} \alpha_{ij} \geq 0$ for $i = 1, \ldots, n$.

Plugging 7 and 8 back into 6, we get

$$
\begin{aligned}
\mathcal{L}(\alpha) &= \frac{1}{2}\vec{w}^\top \vec{w} + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \sum_{z_j \in \mathbb{K}_i} \alpha_{ij}(\vec{w}^\top \Psi_i(z_j) - \Delta_i(z_j) + \xi_i) - \sum_{i=1}^{n} \beta_i \xi_i \\
&= \frac{1}{2}\vec{w}^\top \vec{w} + \sum_{i=1}^{n} \xi_i \left(C - \beta_i - \sum_{z_j \in \mathbb{K}_i} \alpha_{ij} \right) - \sum_{i=1}^{n} \sum_{z_j \in \mathbb{K}_i} \alpha_{ij}(\vec{w}^\top \Psi_i(z_j) - \Delta_i(z_j)) \\
&= \frac{1}{2}\vec{w}^\top \vec{w} - \sum_{i=1}^{n} \sum_{z_j \in \mathbb{K}_i} \alpha_{ij}(\vec{w}^\top \Psi_i(z_j) - \Delta_i(z_j)) \\
&= \frac{1}{2}\vec{w}^\top \vec{w} - \vec{w}^\top \sum_{i=1}^{n} \sum_{z_j \in \mathbb{K}_i} \alpha_{ij} \Psi_i(z_j) + \sum_{i=1}^{n} \sum_{z_j \in \mathbb{K}_i} \alpha_{ij} \Delta_i(z_j) \\
&= -\frac{1}{2}\vec{w}^\top \vec{w} + \sum_{i=1}^{n} \sum_{z_j \in \mathbb{K}_i} \alpha_{ij} \Delta_i(z_j) \\
&= -\frac{1}{2} \left\| \sum_{i=1}^{n} \sum_{z_j \in \mathbb{K}_i} \alpha_{ij} \Psi_i(z_j) \right\|^2 + \sum_{i=1}^{n} \sum_{z_j \in \mathbb{K}_i} \alpha_{ij} \Delta_i(z_j).
\end{aligned}
$$

Thus, the dual problem is

$$
\begin{aligned}
&\underset{\alpha}{\text{maximize}} \;\; -\frac{1}{2} \left\| \sum_{i=1}^{n} \sum_{z_j \in \mathbb{K}_i} \alpha_{ij} \Psi_i(z_j) \right\|^2 + \sum_{i=1}^{n} \sum_{z_j \in \mathbb{K}_i} \alpha_{ij} \Delta_i(z_j) \\
&\text{subject to } 0 \leq \sum_{z_j \in \mathbb{K}_i} \alpha_{ij} \leq C \\
&\qquad\qquad\quad 0 \leq \alpha_{ij}, \qquad \forall i, \forall j.
\end{aligned}
\tag{9}
$$

Bonus: We note that in the dual form, constraints are separable in blocks which is favorable for optimization (see *https://arxiv.org/pdf/1207.4747.pdf* for more details)