Problem Set #2: Kernels, SVMs, and Theory

Trung H. Nguyen

1. **Kernel ridge regression**

   (a) We have:

$$
\begin{aligned}
J(\theta) &= \frac{1}{2}\sum_{i=1}^{m}\left(\theta^T x^{(i)} - y^{(i)}\right)^2 + \frac{\lambda}{2}\|\theta\|^2 \\
&= \frac{1}{2}\sum_{i=1}^{m}\left(\theta^T x^{(i)} - y^{(i)}\right)\left(\theta^T x^{(i)} - y^{(i)}\right) + \frac{\lambda}{2}\theta^T\theta \\
&= \frac{1}{2}\begin{bmatrix}\theta^T x^{(1)} - y^{(1)} \\ \vdots \\ \theta^T x^{(m)} - y^{(m)}\end{bmatrix}^T \begin{bmatrix}\theta^T x^{(1)} - y^{(1)} \\ \vdots \\ \theta^T x^{(m)} - y^{(m)}\end{bmatrix} + \frac{\lambda}{2}\theta^T\theta \\
&= \frac{1}{2}(X\theta - y)^T(X\theta - y) + \frac{\lambda}{2}\theta^T\theta
\end{aligned}
$$

   Hence, the derivative of $J(\theta)$ is:

$$\nabla_\theta J(\theta) = X^T X\theta - X^T y + \lambda\theta$$

   Setting the gradient to zero, we have:

$$\theta = (X^T X + \lambda I)^{-1} X^T y$$

   (b) Let Φ denote the design matrix corresponding to $\phi(x^{(i)})$'s, or:

$$\Phi = \begin{bmatrix}\phi(x^{(1)})^T \\ \vdots \\ \phi(x^{(m)})^T\end{bmatrix}$$

   Using the result obtained from part (a), we have:

$$
\begin{aligned}
\theta &= (\Phi^T\Phi + \lambda I)^{-1}\Phi^T y \\
&= \Phi^T(\Phi\Phi^T + \lambda I)^{-1} y
\end{aligned}
$$

   Let K be the kernel matrix, then $K\left((\phi(x^{(i)}), \phi(x^{(j)})\right) = K_{ij} = \phi(x^{(i)})^T\phi(x^{(j)})$, which makes $K = \Phi\Phi^T$ and it can be easily seen that K is symmetric. We have:

$$
\begin{aligned}
y_{\text{new}} &= \theta^T\phi(x_{\text{new}}) \\
&= y^T(K + \lambda I)^{-T}\Phi\phi(x_{\text{new}}) \\
&= y^T(K + \lambda I)^{-1}\begin{bmatrix}\phi(x^{(1)})^T\phi(x_{\text{new}}) \\ \vdots \\ \phi(x^{(m)})^T\phi(x_{\text{new}})\end{bmatrix} \\
&= \sum_{i=1}^{m}\left((K + \lambda)^{-1}y\right)_i K\left(\phi(x^{(i)}), \phi(x_{\text{new}})\right) \\
&= \sum_{i=1}^{m}\alpha_i K\left(\phi(x^{(i)}), \phi(x_{\text{new}})\right)
\end{aligned}
$$

2. ℓ_2 **norm soft margin SVMs**

$$\begin{aligned}\min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + \frac{C}{2}\textstyle\sum_{i=1}^m \xi_i^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i,\ i = 1, ..., m\end{aligned}$$

(a) Assume that there is an solution to the problem with some $\xi_i < 0$. Then the constraint $y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i$ is also satisfied with $\xi_i = 0$, and also makes our objective function lower, which will prove that $\xi_i < 0$ is not an optimal solution any more.

(b) The Lagrangian of the ℓ_2 soft margin SVM optimization problem is:

$$\mathcal{L}(w, b, \xi, \alpha) = \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^m \xi_i^2 - \sum_{i=1}^m \alpha_i \left(y^{(i)}(w^T x^{(i)} + b) - 1 + \xi_i\right) \tag{1}$$

(c) Taking the gradient of $\mathcal{L}$ w.r.t w, b, ξ respectively, we have:

$$\begin{aligned}\nabla_w \mathcal{L} &= w - \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} \\ \frac{\partial \mathcal{L}}{\partial b} &= -\sum_{i=1}^m \alpha_i y^{(i)} \\ \nabla_\xi \mathcal{L} &= C\xi - \alpha\end{aligned}$$

Setting each gradient equal to 0, we have:

$$\begin{aligned}w &= \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} \\ \sum_{i=1}^m \alpha_i y^{(i)} &= 0 \\ \alpha_i &= C\xi_i,\ \text{for } i = 1, ...m\end{aligned}$$

(d) Plug the results obtained in part (c) back into the Lagrangian (equation (1)), we have:

$$\mathcal{L}(w, b, \xi, \alpha) = -\frac{1}{2}\sum_{i,j=1}^m \alpha_i \alpha_j y^{(i)} y^{(j)} {x^{(i)}}^T x^{(j)} - \frac{1}{2C}\sum_{i=1}^m \alpha_i^2 + \sum_{i=1}^m \alpha_i$$

Putting this together with the constraints $\alpha_i \geq 0$ and the constraint $\sum_{i=1}^m \alpha_i y^{(i)} = 0$, we obtain the dual formulation of the ℓ_2 soft norm SVM optimization problem:

$$\begin{aligned}\max_\alpha \quad & W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2C}\sum_{i=1}^m {\alpha_i}^2 - \frac{1}{2}\sum_{i,j=1}^m \alpha_i \alpha_j y^{(i)} y^{(j)} {x^{(i)}}^T x^{(j)} \\ \text{s.t} \quad & \alpha_i \geq 0,\ i = 1, ..., m \\ & \sum_{i=1}^m \alpha_i y^{(i)} = 0\end{aligned}$$

3. **SVM with Gaussian kernel**

(a) Let $\alpha_i = 1$ for all i and $b = 0$, combine with the fact that $y \in \{-1; 1\}$, for a training example

$(x^{(i)}, y^{(i)})$, we have:

$$\begin{aligned}
\left|f(x^{(i)}) - y^{(i)}\right| &= \left|\sum_{j=1}^{m} y^{(j)} K(x^{(j)}, x^{(i)}) - y^{(i)}\right| \\
&= \left|\sum_{j=1, j\neq i}^{m} y^{(j)} \exp\left(-\frac{\left\|x^{(j)} - x^{(i)}\right\|^2}{\tau^2}\right)\right| \\
&\leq \sum_{j=1, j\neq i}^{m} \left|y^{(j)} \exp\left(-\frac{\left\|x^{(j)} - x^{(i)}\right\|^2}{\tau^2}\right)\right| \\
&\leq \sum_{j=1, j\neq i}^{m} \exp\left(-\frac{\epsilon^2}{\tau^2}\right) \\
&= (m-1) \exp\left(-\frac{\epsilon^2}{\tau^2}\right)
\end{aligned}$$

To make $\left|f(x^{(i)}) - y^{(i)}\right| < 1$ for all i, we need to choose a value of τ such that:

$$\begin{aligned}
(m-1) \exp\left(-\frac{\epsilon^2}{\tau^2}\right) &< 1 \\
\frac{\epsilon^2}{\tau^2} &> \log(m-1) \\
\tau &< \frac{\epsilon}{\sqrt{\log(m-1)}}
\end{aligned}$$

(b)

4. **Naive Bayes and SVMs for Spam Classification**

5. **Uniform convergence**

(a) Let any γ be fixed such that $1 > \gamma > 0$, and let $h \in \mathcal{H}$ be a hypothesis with $\varepsilon(h) > \gamma$. We have for any $i \in [1, m]$:

$$\begin{aligned}
p\left(h(x^{(i)}) \neq y^{(i)}\right) &> \gamma \\
p\left(h(x^{(i)}) = y^{(i)}\right) &\leq 1 - \gamma
\end{aligned}$$

Therefore, using the assumption that training examples are drawn independently, we have:

$$\begin{aligned}
p\left(h(x^{(i)}) = y^{(i)} \quad \forall i = 1, ..., m\right) &\leq (1-\gamma)^m \\
&\leq e^{-\gamma m}
\end{aligned}$$

Thus, using the union bound, we have that:

$$\begin{aligned}
p(\exists h \in \mathcal{H}. \varepsilon(h) > \gamma; h(x^{(i)}) = y^{(i)} \quad \forall i = 1, ..., m) &\leq k e^{-\gamma m} \\
p(\forall h \in \mathcal{H}. \varepsilon(h) \leq \gamma; h(x^{(i)}) = y^{(i)} \quad \forall i = 1, ..., m) &\geq 1 - k e^{-\gamma m}
\end{aligned}$$

Let $\delta = k e^{-\gamma m}$, which give us:

$$\gamma = \frac{1}{m} \log \frac{k}{\delta}$$

Wrapping everything up, we have with the probability at least $1 - \delta$:

$$\varepsilon(\hat{h}) \leq \frac{1}{m} \log \frac{k}{\delta}$$

(b) For fixed δ, γ, for $\varepsilon(\hat{h}) \leq \gamma$ to hold with probability at least $1 - \delta$, it suffices that:

$$m \geq \frac{1}{\gamma} \log \frac{k}{\delta}$$