1.Kernel ridge regression

(a)记

$$X = \begin{bmatrix} (x^{(1)})^T \\ (x^{(2)})^T \\ \dots \\ (x^{(m)})^T \end{bmatrix}, \vec{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \dots \\ y^{(m)} \end{bmatrix}$$

所以

$$\begin{aligned} J(\theta) &= \frac{1}{2}(X\theta - \vec{y})^T(X\theta - \vec{y}) + \frac{\lambda}{2}\theta^T\theta \\ &= \frac{1}{2}(\theta^T X^T X\theta - 2\vec{y}^T X\theta + \vec{y}^T\vec{y}) + \frac{\lambda}{2}\theta^T\theta \end{aligned}$$

关于θ求梯度可得

$$\begin{aligned} \nabla_\theta J(\theta) &= \frac{1}{2}(2X^T X\theta - 2X^T\vec{y}) + \lambda\theta \\ &= (\lambda I + X^T X)\theta - X^T\vec{y} \end{aligned}$$

令上式为0可得

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

(b)首先证明题目中的等式：

$$(\lambda I + BA)^{-1}B = B(\lambda I + AB)^{-1}$$

因为

$$B(\lambda I + AB) = B + BAB = (\lambda I + BA)B$$

所以

$$(\lambda I + BA)^{-1}B = B(\lambda I + AB)^{-1}$$

记

$$\tilde{X} = \begin{bmatrix} (\phi(x^{(1)}))^T \\ (\phi(x^{(2)}))^T \\ \dots \\ (\phi(x^{(m)}))^T \end{bmatrix}$$

所以由(a)可得

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

从而

$$\theta^T\phi(x_{\rm new})=\vec y^T\tilde X(\lambda I+\tilde X^T\tilde X)^{-1}\phi(x_{\rm new})$$

对等式

$$(\lambda I+BA)^{-1}B=B(\lambda I+AB)^{-1}$$

取

$$A=\tilde X^T,B=\tilde X$$

可得

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

带回原式得到

$$\theta^T\phi(x_{\rm new})=\vec y^T(\lambda I+\tilde X\tilde X^T)^{-1}\tilde X\phi(x_{\rm new})$$

下面分别计算$\tilde X\tilde X^T,\tilde X\phi(x_{\rm new})$:

$$\begin{aligned}
\tilde X\tilde X^T&=\left[\begin{matrix}(\phi(x^{(1)}))^T\\(\phi(x^{(2)}))^T\\\ldots\\(\phi(x^{(m)}))^T\end{matrix}\right]\left[\begin{matrix}\phi(x^{(1)}) & \phi(x^{(2)}) & \ldots & \phi(x^{(m)})\end{matrix}\right]\\
&=[\phi(x^{(i)})^T\phi(x^{(j)})]_{i,j}\\
\tilde X\phi(x_{\rm new})&=\left[\begin{matrix}(\phi(x^{(1)}))^T\\(\phi(x^{(2)}))^T\\\ldots\\(\phi(x^{(m)}))^T\end{matrix}\right]\phi(x_{\rm new})\\
&=\left[\begin{matrix}(\phi(x^{(1)}))^T\phi(x_{\rm new})\\(\phi(x^{(2)}))^T\phi(x_{\rm new})\\\ldots\\(\phi(x^{(m)}))^T\phi(x_{\rm new})\end{matrix}\right]
\end{aligned}$$

所以每一项只与内积有关，不需要计算$\phi(x_{\rm new})$

2.ℓ_2 norm soft margin SVMs

(a)只要说明最优解必然满足$\xi_i\ge 0,\forall i=1,\ldots,m$即可，利用反证法，假设存在$\xi_j<0$，那么

$$y^{(j)}(w^Tx^{(j)}+b)\ge 1-\xi_j>1$$

此时目标函数为

$$\frac{1}{2}||w||^2 + \frac{C}{2}\sum_{i=1}^{m}\xi_i^2 \tag{1}$$

现在取$\xi'_j = -\frac{\xi_j}{2} > 0$，那么

$$y^{(j)}(w^Tx^{(j)} + b) \ge 1 - \xi_j > 1 > 1 - \xi'_j = 1 + \frac{\xi_j}{2}$$

但是此时目标函数为

$$\frac{1}{2}||w||^2 + \frac{C}{2}\sum_{i\neq j}\xi_i^2 + \frac{C}{2}\xi_j'^2 = \frac{1}{2}||w||^2 + \frac{C}{2}\sum_{i\neq j}\xi_i^2 + \frac{C}{8}\xi_j^2 \tag{2}$$

(1)减去(2)可得

$$\frac{3C}{8}\xi_j^2 > 0$$

这就与(1)是最小值矛盾，从而原假设成立。

(b)优化问题为

$$\begin{aligned}\min_{\gamma,w,b}\ & \frac{1}{2}||w||^2 + \frac{C}{2}\sum_{i=1}^{m}\xi_i^2 \\ \text{s.t}\ & y^{(i)}(w^Tx^{(i)} + b) \ge 1 - \xi_i, i = 1,\ldots,m\end{aligned}$$

将条件化为标准形式

$$1 - \xi_i - y^{(i)}(w^Tx^{(i)} + b) \le 0, i = 1,\ldots,m$$

我们可以得到拉格朗日算子

$$\mathcal{L}(w,\beta,\xi,\alpha) = \frac{1}{2}||w||^2 + \frac{C}{2}\sum_{i=1}^{m}\xi_i^2 - \sum_{i=1}^{m}\alpha_i[y^{(i)}(w^Tx^{(i)} + b) - 1 + \xi_i] \tag{3}$$

这里，α_i拉格朗日乘子（约束为≥ 0）。

(c)求偏导并令为0可得：

$$\begin{aligned}\nabla_w\mathcal{L}(w,\beta,\xi,\alpha) &= w - \sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)} = 0 \\ \nabla_b\mathcal{L}(w,\beta,\xi,\alpha) &= \sum_{i=1}^{m}\alpha_i y^{(i)} = 0 \\ \nabla_{\xi_i}\mathcal{L}(w,\beta,\xi,\alpha) &= C\xi_i - \alpha_i = 0\end{aligned}$$

化简得到

$$w=\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)} \tag{4}$$

$$\sum_{i=1}^{m}\alpha_i y^{(i)}=0 \tag{5}$$

$$C\xi_i=\alpha_i \tag{6}$$

(d)将等式(4)带入(3)可得

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

将等式(5)带入可得

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

将等式(6)带入可得

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

所以对偶问题为

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

3.SVM with Gaussian kernel

(a)按提示取

$$\alpha_i=0, i=1,\ldots,m, b=0$$

因为$y\in\{-1,+1\}$，所以当下式满足时

$$|f(x^{(j)})-y^{(j)}|<1$$

$f(x^{(j)})$与$y^{(j)}$同号，即此时预测正确，接下来找到τ使得上述不等式对任意$j=1,\ldots,m$都成立。

首先计算$f(x^{(j)})$

$$f(x^{(j)})=\sum_{i=1}^{m}y^{(i)}K(x^{(i)},x^{(j)})$$

注意到

$$K(x,x)=1$$

那么

$$\begin{aligned}f(x^{(j)})-y^{(j)}&=\sum_{i\neq j}y^{(i)}K(x^{(i)},x^{(j)})+y^{(j)}K(x^{(j)},x^{(j)})-y^{(j)}\\&=\sum_{i\neq j}y^{(i)}K(x^{(i)},x^{(j)})\end{aligned}$$

现在考虑$|f(x^{(j)})-y^{(j)}|$的上界：

$$\begin{aligned}|f(x^{(j)})-y^{(j)}|&=|\sum_{i\neq j}y^{(i)}K(x^{(i)},x^{(j)})|\\&\le\sum_{i\neq j}|y^{(i)}K(x^{(i)},x^{(j)})|\\&\le\sum_{i\neq j}K(x^{(i)},x^{(j)})\end{aligned}$$

注意到条件有$||x^{(j)}-x^{(i)}||>\epsilon$，所以

$$K(x^{(i)},x^{(j)})=\exp(-||x^{(j)}-x^{(i)}||^2/\tau^2)\le\exp(-\epsilon^2/\tau^2)$$

因此

$$|f(x^{(j)})-y^{(j)}|\le\sum_{i\neq j}K(x^{(i)},x^{(j)})\le(m-1)\exp(-\epsilon^2/\tau^2)$$

如果我们有

$$(m-1)\exp(-\epsilon^2/\tau^2)<1$$

那么对于任意j，必然有

$$|f(x^{(j)})-y^{(j)}|<1$$

求解不等式可得

$$\begin{aligned}m-1&<\exp(\epsilon^2/\tau^2)\\\log(m-1)&\le\epsilon^2/\tau^2\\\tau&<\frac{\epsilon}{\log(m-1)}\end{aligned}$$

所以只要满足上述不等式即可。

（备注，题目中取$\alpha_i=1$，但是实际应该满足

$$\sum_{i=1}^{m}\alpha_i y^{(i)}=0$$

由于$y^{(i)} \in \{-1,+1\}$，所以总存在$M>0$，使得$|\alpha_i|<M$且

$$\sum_{i=1}^m \alpha_i y^{(i)} = 0$$

在这个条件下，对之前的不等式稍作修改即可，结论依然成立。）

(b)由(a)可知存在w，使得样本分类正确，即

$$y^{(i)}(w^T x^{(i)}+b)>0, i=1,\ldots,m \tag{1}$$

由(a)可知取$b=0$，那么(1)化为

$$y^{(i)}(w^T x^{(i)})>0, i=1,\ldots,m$$

注意到

$$w=\sum_{i=1}^m \alpha_i y^{(i)} x^{(i)}$$

让α_i乘以一定的倍数，必然可以使下式

$$y^{(i)}(w^T x^{(i)}) \ge 1, i=1,\ldots,m$$

从而训练误差为0。

(c)不一定，因为我们在最小化

$$\frac{1}{2}||w||^2 + C\sum_{i=1}^m \xi_i$$

如果C很小，那么$C\sum_{i=1}^m \xi_i$的值很小，所以使上式最小的参数可能存在$\xi_i>0$，即训练误差不等于0。

4.Naive Bayes and SVMs for Spam Classification

见2017版的作业。

5.Uniform convergence

(a)首先回顾定义：

$$\hat\epsilon(h) = \frac{1}{m}\sum_{i=1}^m 1\{h(x^{(i)}) \ne y^{(i)}\}$$
$$\epsilon(h) = P_{(x,y)\sim\mathcal{D}}(h(x)\ne y)$$

考虑如下概率：

$$
\begin{aligned}
P(|\epsilon(h)-\hat{\epsilon}(h)|>\gamma, \hat{\epsilon}(h)=0) &= P(|\epsilon(h)-\hat{\epsilon}(h)|>\gamma|\hat{\epsilon}(h)=0)P(\hat{\epsilon}(h)=0)\\
&= P(|\epsilon(h)|>\gamma)\prod_{i=1}^{m}P(1\{h(x^{(i)})\neq y^{(i)}\})\\
&= P(\epsilon(h)>\gamma)(1-\epsilon(h))^m\\
&= 1\{\epsilon(h)>\gamma\}(1-\epsilon(h))^m\\
&\le (1-\gamma)^m\\
&\le e^{-\gamma m}
\end{aligned}
$$

A_i表示事件$|\epsilon(h_i)-\hat{\epsilon}(h_i)|>\gamma, \hat{\epsilon}(h_i)=0$，注意题目的条件为存在$h$，使得$\hat{\epsilon}(h)=0$，所以

$$
\exists h\in\mathcal{H}, |\epsilon(h)-\hat{\epsilon}(h)|>\gamma
$$

与

$$
A_1\bigcup\ldots\bigcup A_k
$$

等价，所以

$$
\begin{aligned}
P(\exists h\in\mathcal{H}, |\epsilon(h)-\hat{\epsilon}(h)|>\gamma) &= P(A_1\bigcup\ldots\bigcup A_k)\\
&\le \sum_{i=1}^{k}P(A_i)\\
&\le \sum_{i=1}^{k}e^{-\gamma m}\\
&\le ke^{-\gamma m}
\end{aligned}
$$

两边同时减1可得

$$
\begin{aligned}
P(\neg\exists h\in\mathcal{H}, |\epsilon(h)-\hat{\epsilon}(h)|>\gamma) &= P(\forall h\in\mathcal{H}. |\epsilon(h)-\hat{\epsilon}(h)|\le\gamma)\\
&\ge 1-ke^{-\gamma m}
\end{aligned}
$$

令$\delta=ke^{-\gamma m}$可得

$$
\begin{aligned}
e^{\gamma m} &= \frac{k}{\delta}\\
\gamma &= \frac{1}{m}\log\frac{k}{\delta}
\end{aligned}
$$

注意$\hat{h}=\arg\min_{h\in\mathcal{H}}\hat{\epsilon}(h)$（此处有$\hat{\epsilon}(\hat{h})=0$），所以有$1-\delta$的概率，如下事件发生

$$
\epsilon(\hat{h})\le\hat{\epsilon}(\hat{h})+\frac{1}{m}\log\frac{k}{\delta}=\frac{1}{m}\log\frac{k}{\delta}
$$

(b)令

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

那么此时

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

解得

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

从而

$$f(k,\gamma,\delta) = \frac{1}{\gamma}\log\frac{k}{\delta}$$