CS 189 Introduction to Machine Learning
Fall 2017 HW11

This homework is due **Monday, November 13 at 10pm.**

1 Getting Started

You may typeset your homework in latex or submit neatly handwritten and scanned solutions. Please make sure to start each question on a new page, as grading (with Gradescope) is much easier that way! Deliverables:

1. Submit a PDF of your writeup to assignment on Gradescope, "HW[n] Write-Up"
2. Submit all code needed to reproduce your results, "HW[n] Code".
3. Submit your test set evaluation results, "HW[n] Test Set".

After you've submitted your homework, be sure to watch out for the self-grade form.

(a) Before you start your homework, write down your team. Who else did you work with on this homework? List names and email addresses. In case of course events, just describe the group. How did you work on this homework? Any comments about the homework?

None - Alone
Comment: N/A

(b) Please copy the following statement and sign next to it:

I certify that all solutions are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in this write up.

I certify that all solns. are entirely in my words & that I have not looked another Stud's sols I have credited all external sources in this writeup

Problem # 2 :

a.) $\langle f, g\rangle_{\mathcal{H}} = \langle g, f\rangle_{\mathcal{H}}$

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{m=1}^{M}\sum_{s=1}^{S} \alpha_m \beta_s k(y_m, x_s)$$

$$= \sum_{s=1}^{S}\sum_{m=1}^{M} \beta_s \alpha_m k(x_s, y_m) = \langle g, f\rangle_{\mathcal{H}}$$

$\langle af, g\rangle_{\mathcal{H}} = a\langle f, g\rangle_{\mathcal{H}}$

$$f = \sum_{m=1}^{M} \alpha_m k(x, y_m) \Rightarrow af = \sum_{m=1}^{M} \underbrace{a\alpha_m}_{\alpha'_m} k(x, y_m)$$

$$\Rightarrow \langle af, g\rangle_{\mathcal{H}} = \sum_{m=1}^{M}\sum_{s=1}^{S} \alpha'_m \beta_s k(y_m, x_s)$$

$$= \sum_{m=1}^{M}\sum_{s=1}^{S} a\alpha_m \beta_s k(y_m, x_s)$$

$$= a\sum_{m=1}^{M}\sum_{s=1}^{S} \alpha_m \beta_s k(y_m, x_s)$$

$$= a\langle f, g\rangle_{\mathcal{H}}$$

$\langle f+g, g\rangle_{\mathcal{H}} = \langle f, g\rangle_{\mathcal{H}} + \langle h, g\rangle_{\mathcal{H}}$

$$f+g = \sum_{m=1}^{M} \alpha_m k(x, y_m) + \sum_{n=1}^{N} \gamma_n k(x, z_n)$$

$$= \sum_{i=1}^{M+N} \delta_i k(x, w_i) \quad \text{where} \begin{cases} \delta_i = \alpha_i & \text{if } i \le M \\ \delta_i = \gamma_{i-M} & \text{if } i > M \end{cases}$$

$$\begin{cases} w_i = y_i & \text{if } i \le M \\ w_i = z_{i-M} & \text{if } i > M \end{cases}$$

$$\Rightarrow \langle f+g, h\rangle_{\mathcal{H}} = \sum_{i=1}^{M+N}\sum_{s=1}^{S} \delta_i \beta_s k(w_i, x_s)$$

$$\sum_{s=1}^{S} \beta_s \left(\sum_{m=1}^{M} \alpha_m k(y_m, x_s) + \sum_{n=1}^{N} \gamma_n k(z_n, x_s) \right)$$

$$= \sum_{s=1}^{S} \sum_{m=1}^{M} \alpha_m \beta_s k(y_m, x_s) + \sum_{s=1}^{S} \sum_{n=1}^{N} \gamma_n \beta_s (z_n, x_s)$$

$$= \langle f, g \rangle_{\mathcal{H}} + \langle h, g \rangle_{\mathcal{H}}$$

$$\langle f, f \rangle_{\mathcal{H}} \geq 0$$

$$\langle f, f \rangle = \sum_{i=1}^{M} \sum_{j=1}^{M} \alpha_i \alpha_j k(y_i, y_j)$$

$$= [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_m] \ K \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix} = \alpha^T K \alpha$$

$$K > 0 \Rightarrow \alpha^T K \alpha > 0 \quad \forall \ \alpha^T \neq 0^T$$

$$\alpha^T K \alpha = 0 \quad \text{iff} \quad \alpha^T = 0^T$$

i.e $\langle f, f \rangle = 0$ iff $f = 0$

The norm of the function f is

$$\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}} = \sqrt{\alpha^T K \alpha}$$

(b) $\langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}} = k(x, y)$

let $f(\cdot) = \sum_{m=1}^{M} \alpha_m k(\cdot, x) = k(\cdot, x) \quad \left(\begin{matrix} \alpha_m = 1 \\ M = 1 \end{matrix} \right)$

$g(\cdot) = \sum_{m=1}^{S} \beta_s k(\cdot, y) = k(\cdot, y) \quad \left(\begin{matrix} \beta_s = 1 \\ S = 1 \end{matrix} \right)$

$$\Rightarrow \langle f, g \rangle = \langle k(\cdot, x), k(\cdot, y) \rangle = k(x, y)$$

$$= \langle k(x, \cdot), k(y, \cdot) \rangle$$

$$\langle k(\cdot, x_i), f \rangle_H = \langle k(\cdot, x_i), \sum_{m=1}^{M} \alpha_m k(x, y_m) \rangle$$

$$= \sum_{m=1}^{M} \alpha_m \underbrace{\langle k(\cdot, x_i), k(x, y_m) \rangle}_{k(x_i, y_m)}$$

$$= \sum_{m=1}^{M} \alpha_m k(x_i, y_m) = f(x_i)$$

(c)

$$\min_{f \in H} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda \|f\|_H^2$$

We define $M =: \{ \sum_{n=1}^{N} \alpha_n k(x, x_n) \}$ to be the subspace of interest, that is the sub space that if $h \in M$,

$L(y_i, h(x_i)) = 0$ for $i = 1 .. n$

We can alway find this subspace given $K > 0$. why?

Because given n value $t_1, t_2 \ldots t_n$ s.t

$L(y_i, t_i) = 0$

we can solve the system of equation

$h(x_i) = t_i$

$\sum_{j=1}^{N} \alpha_j k(x_i, x_j) = t_i$

$K\alpha = t \quad (*)$

since K is invertible $\Rightarrow$ $(*)$ always has solution α

α exists $\rightarrow$ M exists

we write: $f = m + g$

The objective function becomes:

$$\min_{f\in\mathcal{H}} F = \min_{f\in\mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, m+g) + \lambda \underbrace{\| m+g \|_{\mathcal{H}}^2}$$

$$\langle m+g, m+g \rangle$$
$$= \langle m,m \rangle + \cancel{\langle m,g \rangle}^{0} + \cancel{\langle g,m \rangle}^{0} + \langle g,g \rangle$$
$$= \langle m,m \rangle + \langle g,g \rangle$$

$$L(y_i, m+g) \geqslant L(y_i, m) = 0$$

$$\Rightarrow \quad F \geqslant \frac{1}{N} \sum_{i=1}^{N} L(y_i, m) + \langle m,m \rangle + \langle g,g \rangle$$

$$\geqslant \frac{1}{N} \sum_{i=1}^{N} L(y_i, m) + \langle m,m \rangle$$

$$\min F = \frac{1}{N} \sum_{i=1}^{N} L(y_i, m) + \langle m,m \rangle$$

iff $f = m$, i.e $f \in M$

thus $f(x) = \sum_{i=1}^{N} \alpha_i k(x, x_i)$

(d) $L(y, f(x)) = \max(0, 1 - y f(x))$

$$\min_{f \in H} \frac{1}{N} \sum_{i=1}^{N} \max(0, 1 - y_i f(x_i)) + \lambda \underbrace{\|f\|_H^2}_{\alpha^T K \alpha}$$

from (c): $f(x) = \sum_{i=1}^{N} \alpha_i k(x, x_i)$

again, we can alway find subspace $M = \{ \sum_{n=1}^{N} \alpha_i k \, k(x, x_n) : \alpha_i \in R \}$

$\Rightarrow$ kernel SVM:

$$\min_{\alpha \in R^d} \frac{1}{N} \sum_{i=1}^{N} \max(0, 1 - y_i \sum_{j=1}^{N} \alpha_j k(x_i, x_j)) + \lambda \alpha^T K \alpha$$

(e) General optimization problem:

$$\min_{f \in H} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda \|f\|_H^2 \quad (*)$$

$$L(y_i, f(x_i)) = (y_i - f(x_i))$$

$$(*) \Rightarrow \min_{f \in H} \frac{1}{N} \underbrace{\sum_{i=1}^{N} (y_i - f(x_i))_2^2}_{\|Y - f(X)\|_2^2} + \lambda \underbrace{\|f\|_H^2}_{\alpha^T K \alpha}$$

$$f(x) = \sum_{i=1}^{N} \alpha_i k(x, x_i)$$

$\Rightarrow f(X) = K\alpha \qquad X = [x_1 \dots x_n]^T$

$$(*) \Rightarrow \min_{\alpha \in R^d} \frac{1}{N} \|Y - K\alpha\|_2^2 + \alpha^T K \alpha$$

let $F = \frac{1}{N}\|Y - K\alpha\|_2^2 + \lambda \alpha^T K \alpha$

$$\frac{dF}{d\alpha} = \frac{d \frac{1}{N}(Y^TY - Y^TK\alpha - \alpha^TK^TY + \alpha^TK^TK\alpha) + \lambda\alpha^TK\alpha}{d\alpha}$$

$$= \frac{1}{N}(0 - 2K^TY + 2K^TK\alpha) + 2\lambda K^T\alpha = 0$$

$$K^TK\alpha - K^TY + 2N\lambda K^T\alpha = 0$$

$$K^T(K\alpha + \lambda N I_N \alpha) = K^TY$$

one solution is $K\alpha + \lambda N I_N \alpha = Y$

$$(K + \lambda N I_N)\alpha = Y$$

$$\alpha = (K + \lambda N I_N)^{-1} Y$$

(f) Consider Tikhonov:

$$\min_w \frac{1}{N}\|Y - Xw\|_2^2 + \Gamma\|w\|_2^2$$

we know that the solution is:

$$w = (X^TX - N\Gamma)^{-1}X^TY$$

we can write:

$$(X^TX - N\Gamma)w = X^TY$$

$$X^TXw - N\Gamma w = X^TY \qquad (*)$$

$$N\Gamma w = X^TXw - X^TY = X^T(Xw - Y)$$

$$w = \frac{1}{N}\Gamma^{-1}X^T(Xw - Y) = \frac{1}{N}\lambda\Gamma^{-1}X^T\underbrace{\frac{(Xw - Y)}{\lambda}}_{a}$$

$$= \frac{1}{N}\lambda\Gamma^{-1}X^Ta$$

Substitute $w = \frac{1}{N} \lambda \Gamma^{-1} X^T \alpha$ into $(*)$

$$(*) \Rightarrow \frac{1}{N} X^T X \lambda \Gamma^{-1} X^T \alpha - \Gamma \lambda \Gamma^{-1} X^T \alpha = X^T Y$$

$$X^T (X \lambda \Gamma^{-1} X^T \alpha - N \lambda I \alpha) = X^T Y$$

one solution is $\quad X \lambda \Gamma^{-1} X^T \alpha - N \lambda I \alpha = Y$

$$\alpha = (X \lambda \Gamma^{-1} X^T - N \lambda I)^{-1} Y$$

where

$$X = \begin{bmatrix} 1 & x_1 & x_2 & x_1 x_2 & x_1^2 & x_2^2 \\ & & \vdots & & & \end{bmatrix}_{N \times 6}$$

Compare to kernel ridge regression

$$\alpha = (K + \lambda N I)^{-1} Y$$

we need

$$X \lambda \Gamma^{-1} X^T = K$$

we know $\quad K(a, b) = [1 \ \sqrt{2} a_1 \ \sqrt{2} a_2 \ \sqrt{2} a_1 a_2 \ a_1^2 a_2^2]^T \begin{bmatrix} 1 \\ \sqrt{2} a_1 \\ \sqrt{2} a_1 \\ \sqrt{2} a_1 a_2 \\ a_1^2 \\ a_2^2 \end{bmatrix}$

(for $len(a) = len(b) = 2$)

$\Rightarrow \quad \lambda \Gamma^{-1}$ must account for coefficient

$$\Rightarrow \quad \lambda \Gamma^{-1} = diag (1, \underbrace{2, \dots, 2}_{\frac{N(N+1)}{2}}, \underbrace{1, \dots 1}_{N})$$

$$\Rightarrow \quad \Gamma = diag (\lambda, \underbrace{\tfrac{\lambda}{2}, \dots, \tfrac{\lambda}{2}}_{N(N+1)/2}, \underbrace{\lambda, \dots, \lambda}_{N})$$

Thus kernel ridge regression is equivalent to a polynomial regression for $d=2$ with Tikhonov regularization

$$\Gamma = \operatorname{diag}(\lambda, \underbrace{\tfrac{\lambda}{2}, \dots, \tfrac{\lambda}{2}}_{\frac{N(N+1)}{2}}, \underbrace{\lambda, \dots, \lambda}_{N})$$

where N is len(sample$_i$)

(g) Use kernel: we have to solve $n \times n$ system of eqns

$\rightarrow$ runtime $O(n^3)$

No use kernel: solve $O(d^p) \times O(d^p)$ system of eqns $\rightarrow$ runtime $O(d^{3p})$.

Problem # 4

Extend problem 3 (f). Find Tikhonov regularization to equate OLS polynomial regression to kernel ridge regression with $d = 3$ and len(sample$_i$) $= 2$.

Solution

$$a = [a_1 \ a_2] \qquad b = [b_1 \ b_2]$$

$$k(a, b) = (1 + a^T b)^3 = (1 + a_1 b_1 + a_2 b_2)^3$$

$$= (1 + 2a_1 b_1 + 2a_2 b_2 + 2a_1 b_1 a_2 b_2 + a_1^2 b_1^2 + a_2^2 b_2^2)(1 + a_1 b_1 + a_2 b_2)$$

$$= 1 + a_1 b_1 + a_2 b_2 + 2a_1 b_1 + 2a_1^2 b_1^2 + 2a_1 b_1 a_2 b_2 + 2a_2 b_2 + 2a_1 b_1 a_2 b_2 + a_2^2 b_2^2$$

$$+ 2a_1 b_1 a_2 b_2 + 2a_1^2 b_1^2 a_1 b_1 + 2a_1 b_1 a_2^2 b_2^2 + a_1^2 b_1^2 + a_1^3 b_1^3 + a_1^2 b_1^2 a_2 b_2$$

$$+ a_2^2 b_2^2 + a_1 b_1 a_2^2 b_2^2 + a_2^3 b_2^3$$

$$= 1 + 3a_1 b_1 + 3a_2 b^2 + 3a_1^2 a_2^2 + 3a_1^2 a_2^2 + 6a_1 b_1 a_2 b_2$$

$$+ 3a_1^2 b_1^2 a_2 b_2 + 3a_1 b_1 a_2^2 b_2^2 + a_1^3 b_1^3 + a_2^3 b_2^3$$

$$\Rightarrow \quad \Gamma = \begin{bmatrix} \lambda & & & & & & & & & 0 \\ & \frac{\lambda}{3} & & & & & & & & \\ & & \frac{\lambda}{3} & & & & & & & \\ & & & \frac{\lambda}{3} & & & & & & \\ & & & & \frac{\lambda}{3} & & & & & \\ & & & & & \frac{\lambda}{6} & & & & \\ & & & & & & \frac{\lambda}{3} & & & \\ & & & & & & & \frac{\lambda}{3} & & \\ & & & & & & & & \frac{\lambda}{1} & \\ 0 & & & & & & & & & \frac{\lambda}{1} \end{bmatrix}$$