1. Kernel SVM

(a) (1) symmetry

$$\langle f, g\rangle_H = \langle g, f\rangle_H$$

$$\langle f, g\rangle_H = \sum_{m=1}^{M}\sum_{s=1}^{S} \alpha_m \beta_s k(y_m, y_s)$$

$$\langle g, f\rangle_H = \sum_{s=1}^{S}\sum_{m=1}^{M} \beta_s \alpha_m k(y_s, y_m)$$

$\because k(y_m, x_s) = k(y_s, y_m)$

$$\therefore \langle g, f\rangle_H = \sum_{s=1}^{S}\sum_{m=1}^{M} \beta_s \alpha_m k(y_m, y_s) = \sum_{m=1}^{M}\sum_{s=1}^{S} \alpha_m \beta_s k(y_m, y_s)$$

$$= \langle f, g\rangle_H$$

(2) linearity $\langle af, g\rangle_H = a\langle f, g\rangle_H$ and $\langle f+h, g\rangle_H = \langle f, g\rangle_H + \langle h, g\rangle_H$

$$\because \langle f, g\rangle_H = \sum_{m=1}^{M}\sum_{s=1}^{S} \alpha_m \beta_s k(y_m, y_s) = \sum_{s=1}^{S} \beta_s \sum_{m=1}^{M} \alpha_m k(y_s, y_m)$$

$$= \sum_{s=1}^{S} \beta_s f(y_s)$$

$$\therefore \langle af, g\rangle_H = \sum_{s=1}^{S} \beta_s a f(y_s) = a \sum_{s=1}^{S} \beta_s f(y_s) = a\langle f, g\rangle_H$$

$$\therefore \langle f+h, g\rangle_H = \sum_{s=1}^{S} \beta_s \left(f(y_s) + h(y_s)\right)$$

$$= \sum_{s=1}^{S} \beta_s f(y_s) + \sum_{s=1}^{S} \beta_s h(y_s)$$

$$= \langle f, g\rangle_H + \langle h, g\rangle_H$$

(3) positive-definiteness

$$\langle f, f\rangle_H = \sum_{m=1}^{M}\sum_{s=1}^{S} \alpha_m \alpha_s k(y_m, y_s) = \alpha^T K \alpha$$

$\because K$ is PSD matrix $\therefore \langle f, f\rangle_H = \alpha^T K \alpha \geq 0$

$$\because f(\cdot) = \sum_{i=1}^{M} \alpha_i k(\cdot, y_m)$$

$$\therefore \langle k(\cdot, x), f\rangle_H = f(x)$$ This is called reproducing property which is proved in part (b)

$$\because \langle k(\cdot, x), k(\cdot, x)\rangle_H = k(x, x)$$

$$\therefore |f(x)|^2 = |\langle k(\cdot, x), f\rangle_H|^2 \leq \langle k(\cdot, x), k(\cdot, x)\rangle_H \cdot \langle f, f\rangle_H$$

$$= k(x, x)\langle f, f\rangle_H$$

$\therefore$ if $\langle f, f\rangle_H = 0$ then $0 \leq |f(x)|^2 \leq k(x, x)\langle f, f\rangle_H = 0$

$\therefore$ $|f(x)|^2 = 0 \Rightarrow f(x) = 0$

$\because$ if $f(x)=0$ $\langle f, f\rangle_{H_K}=0$ is trivial

$\therefore$ $\langle f, f\rangle=0$ if and only if $f$ is an constant zero function.

$$\therefore \|f\|_H = \sqrt{\langle f, f\rangle_H} = \sqrt{\sum_{m=1}^{M}\sum_{s=1}^{S}\alpha_m\alpha_s k(y_m, y_s)}$$

$$= \sqrt{\alpha^T K \alpha}$$

(b) $\langle k(x,\cdot), k(y,\cdot)\rangle_H$

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

$f: \to k(\cdot, x)$ $\quad g: \to k(\cdot, y)$

$\therefore = k(x, y)$

As for: $\langle k(\cdot, x_i), f\rangle_H = f(x_i)$

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

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

$= f(x_i)$

(c) Hint: Define $\mathcal{M} = \{\sum_{n=1}^{N}\alpha_n k(x, x_n) : \alpha_i \in \mathbb{R}\}$ to be the subspace of interest.

Proof: $\mathcal{M} = \{\sum_{i=1}^{N}\alpha_i k(x, x_i) : \alpha_i \in \mathbb{R}\}$ is a subspace of interest

Project $f$ onto this subspace obtaining $f_S$ (the component along the subspace) and $f_\perp$ (the component perpendicular to the subspace)

$$f = f_S + f_\perp \Rightarrow \|f\|_H^2 = \|f_S\|_H^2 + \|f_\perp\|_H^2 \geq \|f_S\|_H^2$$

$\therefore \|f\|_H$ is minimized if $f$ lies in the subspace.

With the reproducing property proved in (b), for each $i$

$$f(x_i) = \langle f, k(\cdot, x_i)\rangle_H = \langle f_S, k(\cdot, x_i)\rangle_H + \langle f_\perp, k(\cdot, x_i)\rangle_H$$

$$= f_S(x_i)$$

$$\therefore \frac{1}{N}\sum_{i=1}^{N} l(f(x_i), y_i) = \frac{1}{N}\sum_{i=1}^{N} l(f_S(x_i), y_i)$$

To minimize the loss, see $\|f_\perp\|_H$ to be 0

$\therefore$ The minimizing solution is : $f(x) = \sum_{i=1}^{N} \alpha_i k(x, x_i)$

(d) $\because f(x_i) = \sum_{j=1}^{N} \alpha_j k(x_i, x_j)$ (given in (c))

$\|f\|_H = \sqrt{\alpha^T K \alpha}$ (calculated in (a))

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

$\Downarrow$

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

2. $L_1$ - Regularization

(a) $J_\lambda(w) = \frac{1}{2}\|y - Xw\|_2^2 + \lambda\|w\|_1$

$= \frac{1}{2}\|y - Xw\|_2^2 + \lambda \sum_{i=1}^{d} |w_i|$

$= \frac{1}{2}(y - Xw)^T(y - Xw) + \lambda \sum_{i=1}^{d} |w_i|$

$= \frac{1}{2}(w^T X^T - y^T)(y - Xw)$ — then use $X_i^T X_i$

$= \frac{1}{2}(w^T X^T y - w^T X^T X w - y^T y + y^T X w)$

$= -\frac{1}{2} y^T y + \frac{1}{2}(w^T X^T y - w^T X^T X w + y^T X w)$

$= g(y) + \frac{1}{2}(w^T X^T y - w^T X^T X w + y^T X w)$

$\therefore J_\lambda(w) = g(y) + \frac{1}{2}(w^T X^T y - w^T X^T X w + y^T X w) + \lambda \sum_{i=1}^{n} |w_i|$

$= g(y) + \frac{1}{2}(w^T X^T y - n w^T w + y^T X w)$

$= g(y) + \sum_{i=1}^{d}\left(\frac{1}{2} w_i^T X_i^T y - \frac{1}{2} n w_i^T w_i + \frac{1}{2} y^T X_i w_i + \lambda|w_i|\right)$

$= g(y) + \sum_{i=1}^{d}\left(\frac{1}{2} X_i^T y w_i - \frac{1}{2} n w_i^2 + \frac{1}{2} y^T X_i w_i + \lambda|w_i|\right)$

$= g(y) + \sum_{i=1}^{d} f(X_i, y, w_i, \lambda)$

$\therefore$ For data with uncorrelated features, one can learn the parameter $w_i$ corresponding to each $i$-th feature independently from other features, one at a time, and get a solution which is equivalent to having learned them all jointly as we normally do.

(b) If $\hat{w}_i > 0$

$J_\lambda(w_i) = \frac{1}{2} g(y) + \frac{1}{2} X_i^T y w_i - \frac{1}{2} n w_i^2 + \frac{1}{2} y^T X_i w_i + \lambda w_i$

$\frac{\partial J_\lambda(w_i)}{\partial w_i} = \frac{1}{2} X_i^T y - n w_i + \frac{1}{2} y^T X_i + \lambda = 0$

$n w_i = X_i^T y + \lambda$

$w_i = \frac{X_i^T y + \lambda}{n}$

(c) If $\hat{w}_i < 0$

$$J_\lambda(w_i) = \frac{1}{d} g(y) + \frac{1}{2} x_i^T y w_i - \frac{1}{2} n w_i^2 + \frac{1}{2} y^T x_i w_i - \lambda w_i$$

$$\frac{\partial J_\lambda(w_i)}{\partial w_i} = \frac{1}{2} x_i^T y - n w_i + \frac{1}{2} y^T x_i - \lambda = 0$$

$$n w_i = x_i^T y - \lambda$$

$$w_i = \frac{x_i^T y - \lambda}{n}$$

(d) When $\hat{w}_i \leq 0$ $\quad x_i^T y + \lambda = 0$ will make $\hat{w}_i = 0$

when $\hat{w}_i \geq 0$ $\quad x_i^T y - \lambda = 0$ will make $\hat{w}_i = 0$

(e) As for ridge regression

$$J_\lambda(w) = \frac{1}{2} \|y - Xw\|_2^2 + \lambda \sum_{i=1}^{d} w_i^2$$

$$= -\frac{1}{2} y^T y + \frac{1}{2} (w^T x^T y - w^T x^T x w + y^T x w) + \lambda \sum_{i=1}^{d} w_i^2$$

$$= g(y) + \sum_{i=1}^{d} \left( \frac{1}{2} w_i x_i^T y - \frac{1}{2} n w_i^2 + \frac{1}{2} y^T x_i w_i + \lambda w_i^2 \right)$$

$$J_\lambda(w_i) = \frac{1}{d} g(y) + \frac{1}{2} x_i^T y w_i - \frac{1}{2} n w_i^2 + \lambda w_i^2 + \frac{1}{2} y^T x_i w_i$$

$$\frac{\partial J_\lambda(w_i)}{\partial w_i} = \frac{1}{2} x_i^T y - n w_i + 2\lambda w_i + \frac{1}{2} y^T x_i$$

$$= (2\lambda - n) w_i + x_i^T y = 0$$

$$w_i = \frac{x_i^T y}{n - 2\lambda}$$

To make $w_i = 0$, $x_i^T y = 0$ which is very hard for training dataset.

But for the result we get in part (d) $x_i^T y + \lambda = 0$ or $x_i^T y - \lambda = 0$ are much easier to satisfy

$\therefore$ $l_1$ - regularization is adapted for feature selection

3. Decision Trees for Classification.

I didn't manage to finish this question