Ashutosh Agarwal (ana63), Yuxiang Peng (yp344),
Yue Sun (ys758), Zhi ting Chen (zc346)

1. (a) $w^* = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$ Since x_is are fixed.

$$E(w^*) = \frac{\sum_{i=1}^n x_i E(y_i)}{\sum_{i=1}^n x_i^2} \quad \because y_i = wx_i + \varepsilon_i$$

$$E(w^*) = \frac{\sum_{i=1}^n x_i E(wx_i + \varepsilon_i)}{\sum_{i=1}^n x_i^2} \qquad \varepsilon_i \sim N(0, \sigma^2)$$

$$\therefore E(w^*) = \frac{\sum_{i=1}^n x_i \cdot wx_i}{\sum_{i=1}^n x_i^2} = w \quad \text{So it's unbiased.}$$

$$Var(w^*) = Var\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\right) = \left(\frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2}\right)^2 \sigma^2 = \left(\frac{\bar{x}}{\overline{x^2}}\right)^2 \sigma^2$$

(b) $w^*(\lambda) = \frac{\sum_{i=1}^n x_i y_i}{\lambda + \sum_{i=1}^n x_i^2} \quad \because \lambda > 0$

It is easy to conclude from (a) $E(w^*(\lambda)) = \frac{w \cdot \sum_{i=1}^n x_i^2}{\lambda + \sum_{i=1}^n x_i^2} \neq w$

$\therefore$ It's ~~not~~ not ~~b~~ unbiased.

$$Var(w^*(\lambda)) = \frac{\bar{x}}{\left(\frac{\lambda}{n} + \overline{x^2}\right)^2} \sigma^2$$

(c) for linear regression

$\bar{h}(x) = E(w^* x) = wx$

$h_D(x) = w^* x$

$\bar{y}(x) = wx + \varepsilon$

$\bar{y}(x) = E(wx + \varepsilon) = wx$

bias $= E_x[(\bar{h}(x) - \bar{y}(x))^2] = 0$

variance $= E_{x,D}[(h_D(x) - \bar{h}(x))^2]$

$= (w^* - w)^2 E(X^2)$

$X \sim \text{unif}(0,1)$ $E(X^2) = \frac{b^2 + ab + a^2}{3} = \frac{1}{3}$

variance $= \frac{1}{3}(w^* - w)^2$

noise $= E_{x,y}[(\bar{y}(x) - y)^2] = \delta^2$

for ridge regression

based on the formula basic from linear regression. The test error $Err(x_0)$ for a ridge regression fit $\hat{f}_\alpha(x_0)$ is identical to the form $Err(x_0) = \sigma^2 + [f(x) - E\hat{f}(x))^2] + \|h(x)\|^2 \sigma^2$

For the variance term, the linear weights is $h(x) = X(X^T X + \alpha I)^{-1} X$

For the bias term, let β_* denote the parameter of the best-fitting linear approximation to f:

$\beta_* = \arg\min_\beta E(f(x) - \beta^T X)^2$. Here the expectation is taken with respect to the distribution of the input variables X. Then the bias is

$E_x[f(x) - E\hat{f}(x)]^2 = E_x[f(x) - \beta_*^T x]^2 + E_x[\beta_*^T x - E\hat{\beta}^T x]^2$

And for the noise: σ^2

(d) If σ^2 is known to be small ahead of time, for the ridge, σ^2 influences the variance term and successfully lowers it by the multiplication with $\|h(x)\|^2$.
Chose the ridge regression. otherwise if σ^2 is large, choose the linear regression model.

Problem 2

(a) $k(x,z) = x^T A z$, where A is positive semidefinite matrix.
It is symmetric, and since A is positive semi-definite,

$$A = B^T B, \quad \text{where } \|B\|^2 \geq 0$$

Thus, $x^T B^T B z$

$$\Rightarrow (Bx)^T (Bz)$$

$$\Rightarrow \|Bx\|^2 \geq 0$$

Thus, $k(x,z)$ is positive semi-definite.
Thus, it is a kernel.

(b) $k(x,z) = c k_1(x,z)$ for some $c \geq 0$

A necessary and sufficient condition is,

$$\int_x \int_z k(x,z)\, g(x)\, g(z)\, dx\, dz \geq 0$$

for all square, integrable functions $g(\cdot)$

Since $k_1(x,z)$ is a valid kernel, (given)

$$\Rightarrow \int_x \int_z k_1(x,z)\, g(x)\, g(z)\, dx\, dz \geq 0$$

$$\Rightarrow c \int_x \int_z k_1(x,z)\, g(x)\, g(z)\, dx\, dz \geq 0 \qquad (c \geq 0)$$

$\Rightarrow k(x,z)$ is a valid kernel.

(c) $k(x,z) = k_1(x,z) + k_2(x,z)$

Since $k_1(x,z)$ and $k_2(x,z)$ are valid kernels,

$$\int_x \int_z k_1(x,z)\, g(x)\, g(z)\, dx\, dz \geq 0 \quad \text{—— (1)}$$

$$\int_x \int_z k_2(x,z)\, g(x)\, g(z)\, dx\, dz \geq 0 \quad \text{—— (2)}$$

From ① and ②, we have,

$$\Rightarrow \underbrace{\int_x \int_z k_1(x,z)\, g(x)\, g(z)\, dx\, dz}_{\geq 0} + \underbrace{\int_x \int_z k_2(x,z)\, g(x)\, g(z)\, dx\, dz}_{\geq 0} \geq 0$$

$\Rightarrow k(x,z)$ is a valid kernel.

(d) $k(x,z) = k_1(x,z)\, k_2(x,z)$

The Gram matrix is given as,

$$K = K_1 \odot K_2 \quad \text{(element-wise product)}$$

$$K_1 = \sum_{i=1}^{n} \lambda_i u_i u_i^T, \quad K_2 = \sum_{j=1}^{n} \mu_j v_j v_j^T, \quad \lambda_i \text{ and } \mu_i \geq 0 \text{ (positive eigen values)}$$

$$K = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \mu_j (u_i u_i^T) \odot (v_j v_j^T)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \mu_j (u_i \odot v_j)(u_i \odot v_j)^T$$

$$= \sum_{l=1}^{n^2} \gamma_l w_l w_l^T$$

Thus, $\forall a \in \mathbb{R}^n$, $a^T K a = \sum_{i=1}^{n^2} \gamma_i a^T w_i w_i^T a$

$$= \sum_{i=1}^{n^2} \gamma_i (w_i^T a)^2 \geq 0$$

Thus, $k(x,z)$ is a valid kernel.

(e) $k(x,z) = q(k_1(x,z))$, where q is a polynomial with non-negative coefficients.

We know that q is a linear combination of powers of the kernel $k_1(x,z)$ with positive coefficients.

From (b), we know that if $k_1(x,z)$ is valid kernel, then, $a\, k_1(x,z)$ is a valid kernel $[a \geq 0]$

$\longrightarrow$ ①

From (d), we know that products of valid kernels are a valid kernel. —— ②

From (c), we know that sums of a valid kernels, are valid kernels themselves —— ③

From ①, ② and ③,
$k(x, z)$ is a valid kernel.

(f) $k(x, z) = e^{k_1(x, z)}$

On expansion, we have,

$$\Rightarrow 1 + k_1(x, z) + \frac{[k_1(x, z)]^2}{2!} + \frac{[k_1(x, z)]^3}{3!} + \dots$$

$$\Rightarrow \lim_{n \to \infty} \sum_{i=0}^{n} \frac{[k_1(x, z)]^i}{i!}$$

Let $\frac{1}{i!} = a_i$. At $\lim_{n \to \infty} a_n = 0$

Thus, all coefficients $\forall i \in [0, \infty)$, a_i is ≥ 0

From (d) we know that products of valid kernels are valid kernels. And from (c), we know that sums are valid.

Thus,
$k(x, z)$ is a valid kernel.

Problem 3

(a) Note that the feature expansion function $\phi: \Sigma^* \to \mathbb{R}^d$ is such that $k(x, z) = \phi(x)^T \cdot \phi(z)$

$G_n(x)$ is the set of n-grams of x

let's suppose the size for $G_n(x)$ is j

and we get $[u_1, u_2 \cdots, u_j]$ for all the substrings

$\because \delta(x, z) = \begin{cases} 1 & \text{if } x = z \\ 0 & \text{otherwise} \end{cases}$

Let's also supose the size for $G_n(z)$ is i.

we get $[s_1, \cdots, s_i]$

$k(x, z) = \sum_{s \in G_n(x)} \sum_{t \in G_n(z)} \delta(s, t)$ means we construct two matrix in the same order. (The same order means the same substring appears in the same location in the two matrixs by applying the function $\phi(x)$). This can be simply done by adding all the substring into one feature set. If the string contains the substring. The value for this location in the matrix is 1, otherwise 0. Now it's obvious it can be done by matrix inner product with 2 independent variable string s and t.

In Math, that is, for string s, t, $|s|$ is the length, $s = s_1 \cdots s_{|s|}$, $s[i:j] = s_i \cdots s_j$, u is a subsequence of s, indices $i = (i_1, \cdots, i_{|u|})$, $l(i) = i_{|u|} - i_1 + 1$. We denote Σ^n: the set of all finite strings of length n, Σ^*: the set of all strings. $\Rightarrow \Sigma^* = \bigcup_{n=0}^{\infty} \Sigma^n$ Then we denote feature sphe $F_n = \mathbb{R}^{\Sigma^n}$. The feature mapping ϕ for a string s is given by defining the u coordinate $\phi_u(s)$ for each $u \in \Sigma^n \Rightarrow \phi_u(s) = \sum_{i: u = s[i]} \lambda^{l(i)}$ (Here $\lambda = 1$). These features measure the number of occurrences of subsequences in a string s weighting them according to their lengths.

$\therefore k(x, z) = \sum_{s \in G_n(x)} \sum_{t \in G_n(z)} \delta(s, t) = \sum_{u \in \Sigma^n} \underbrace{\sum_{i: u = s[i]} \lambda^{l(i)}}_{\phi_u(s)} \underbrace{\sum_{j: u = t[j]} \lambda^{l(j)}}_{\phi_u(t)} = \sum_{u \in \Sigma^n} \langle \phi_u(s) \cdot \phi_u(t) \rangle$

problem 3

b. Base on the proof of question a. we get the feature expansion for string kernels.

$$K(X,Z) = \sum_{s \in G_n(X)} \sum_{t \in G_n(Z)} \delta(s,t) = \sum_{u \in \Sigma^n} \langle \phi_u(s) \cdot \phi_u(t) \rangle$$

In question a, we basically create the sparse matrix (0/1) for the matrix inner product. Let's suppose the size for set X = [illegible], the size for Z = [illegible] m, The feature space $F = \mathbb{R}^{\Sigma^n}$.

$$\therefore K(X,Z) = \sum_{s \in G(X)} \sum_{t \in G(Z)} \delta(s,t) = \sum_{u \in \Sigma^n} \sum$$

The sub-word from a specific set is w. w have two range from 1~[illegible] and from 1~[illegible] m

$$\therefore K(X,Z) = \sum_{u \in \Sigma^n} \sum_{w} \sum_{X[L]} \lambda^{(6)} \cdot \lambda^{l(i)} \cdot \sum_{w \in Z[m]} \sum_{u = t[j]} \lambda^{(m)} \cdot \lambda^{l(j)}$$

Let's denote $\phi^*(X) = \sum_{X[i]} \sum_{S[i]} \lambda^{(l(s))} \cdot \lambda^{(l(i))}$

$$\phi^*(Z) = \sum_{Z[m]} \sum_{t[j]} \lambda^{(m)} \cdot \lambda^{(l(j))}$$

$$\therefore K(X,Z) = \sum_{u \in \Sigma^n} \langle \phi^*(X) \cdot \phi^*(Z) \rangle \text{ (feature expansion)}$$

Question 4

a) $$\nabla L(w) = -\sum_{i=1}^{n} \frac{y_i x_i}{1+\exp(y_i w^T x_i)}$$

$$= \sum_{i=1}^{n} \gamma_i x_i \quad , \quad \gamma_i = \frac{-y_i}{1+\exp(y_i w^T x_i)}$$

b) $$W_{t+1} = W_t - s\frac{\partial l(w)}{\partial w}$$

$$= W_t + s \times \sum_{i=1}^{n} \frac{y_i x_i}{1+\exp(y_i w_t^T x_i)}$$

$$= \sum_{i=1}^{n} \alpha_i^t x_i$$

$$\therefore W = \sum_{i=1}^{n} \alpha_i x_i \qquad W \text{ can be written as linear combination of } X.$$

$$p(y|x_t) = \frac{1}{1+\exp(-y w^T x_t)}$$

$$= \frac{1}{1+\exp(-y\sum_{i=1}^{n}\alpha_i x_i^T x_t)}$$

$$= \frac{1}{1+\exp(-y\sum_{i=1}^{n}\alpha_i k(x_i, x_t))} = \frac{1}{1+\exp(-y\, k(X, x_t)\alpha)}$$

$$\frac{p(y=1|x_t)}{p(y=0|x_t)} = \frac{\frac{1}{1+\exp(-k(X,x_t)\alpha)}}{1/2} = \frac{2}{1+\exp(-k(X,x_t)\alpha)}$$

c) $l(w) = \sum_{i=1}^{n} \log(1 + e^{-y_i w^T x_i})$

$l(\alpha) = \sum_{i=1}^{n} \log(1 + e^{-y_i \sum_{j=1}^{n} \alpha_j x_j^T x_i})$

$= \sum_{i=1}^{n} \log(1 + e^{-y_i \sum_{j=1}^{n} \alpha_j k(x_j, x_i)})$

$= \sum_{i=1}^{n} \log(1 + e^{-y_i K(X, x_i)\alpha})$

$$\frac{\partial l(\alpha)}{\partial \alpha} = \sum_{i=1}^{n} \frac{e^{-y_i k(X, x_i)\alpha}}{1 + e^{-y_i k(X, x_i)\alpha}} \times -y_i k(X, x_i)$$

d) $w = \sum_{j=1}^{n} \alpha_j X_j = X\alpha$

$l(\alpha) = \sum_{i=1}^{n} \log(1 + e^{-y_i \alpha^T X^T x_i}) + \lambda \|X\alpha\|_2^2$

$k(X, x_i) = X^T x_i, \quad k(X, X) = X^T X$

$l(\alpha) = \sum \log(1 + e^{-y_i \alpha^T k(x, x_i)}) + \lambda \underbrace{\alpha^T X^T X \alpha}_{k(x,x)}$

$$= \sum_{i=1}^{n} \frac{-e^{-y_i \alpha^T k(x, x_i)} y_i k(x, x_i)}{1 + e^{-y_i \alpha^T k(x, x_i)}} + \alpha^T (k(x, x) + k^T(x, x))$$