hw6 stat154.

1. computational complexity.

① $a+v$ $O(d)$

a^Tv $O(d)$

② $A+B$. $O(n\cdot d)$.

space storing matrix A. $O(n\cdot d)$

③. Av. $O(n\cdot d^2)$.

A^TB $O(d\cdot n^2)$.

④. A^TBv

2. kernel methods.

① computing X^TX takes ~~O(...)~~ $O(l^2n)$.

inversing matrix takes. $O(l^3)$

thus adding up these leading terms.

computing $\hat{\theta}\lambda$ takes $O(l^3+n l^2)$ complexity.

②. Using raw data matrix X.

original form $w^* = (X^TX+\lambda I)^{-1}X^Ty$.

$(X^TX+\lambda I)w^* = X^Ty$

$X^TXw^*+\lambda w^* = X^Ty$

$\lambda w^* = X^Ty - X^TX\cdot w^*$.

$w^* = \frac{X^T(y-X\cdot w^*)}{\lambda}$. which is the linear combination of columns of X^T.

i.e. $W = X^T\cdot v$

substitute into $X^TXw+\lambda w = X^Ty$

$\Rightarrow X^TX(X^Tv)+\lambda(X^Tv) = X^Ty$

$X^T(XX^Tv+\lambda v) = X^Ty$

we cannot cancel X^T directly, but if $XX^Tv+\lambda v=y$ has solutions.

it implies that $X^T(XX^Tv+\lambda v) = X^Ty$ also has solutions.

$XX^Tv + \lambda v = y$

$v^* = (XX^T+\lambda)^{-1}y$.

substitute v^* back into $w = X^Tv$

$\Rightarrow \quad w = X^T(XX^T+\lambda)^{-1}y$

i.e. $\hat{\theta}_\lambda = X^T(XX^T+\lambda I_n)^{-1}y$ is an equivalent & valid estimate for ridge problem.

complexity: computing XX^T takes $O(n^2 l)$

inverting matrix takes $O(n^3)$

thus adding up those two leading terms. we get $O(n^3+n^2 l)$

③ For $\hat{\theta}_\lambda = (X^TX+\lambda I)^{-1}X^Ty$ .... $O(l^3+l^2 n)$

For $\hat{\theta}_\lambda = X^T(XX^T+\lambda I_n)^{-1}y$ .... $O(n^3+n^2 l)$.

④. ~~we extend~~ the feature map.

A feature map is ~~ap~~ used when the problem cannot be linearly solved. feature map ϕ, map data to a new space where a linear model could be applied to solve the original learning problem.

In addition, kernel function is defined to map ϕ to a higher dimensional space and compute inner-product based similarity $\phi(x)^T\phi(z)$ in that space.

⑤

(a).