Homework-4

Given

Design matrix $= X$

covariance matrix $= C = \frac{1}{n} X^T X$

Eigenvectors: $V_1, V_2 \ldots\ldots V_K$

$\tilde{x}_i$ is the orthogonal projection of x_i, Column vector representing a data point, orthogonal to V_1

$$\tilde{x}_i = (I - V_1 V_1^T) x_i$$

$\tilde{X}_T = [\tilde{x}_1, \ldots\ldots \tilde{x}_n]$ defiated matrix, lies in the direction of the first principal eigenvector

$$\tilde{X}_T = (I - V_1 V_1^T) X^T$$

$$C = \frac{1}{n} X^T X = V \Lambda_n V^T$$

Let's Suppose that the kernel matrix is $K = XX^T$

So, $K = XX^T = U \Lambda_m U^T$

The columns v_i of the orthonormal matrix V are the eigenvectors of C and the columns u_i of the orthonorm

Scanned with CamScanner

matrix U are the eigenvectors of K. Considering an eigen-vector eigen-value pair u, λ of K,

$$C = \frac{1}{n}(X^T u) = X^T X X^T u = X^T K u = \lambda X^T u$$

$X^T u, \lambda$ is an eigenvector-eigenvalue pair for C.

$$\|X^T u\|^2 = u^T X X^T u = \lambda$$

So the normalised eigenvector of C is $v_1 = \lambda^{-\frac{1}{2}} X^T u$.

$$u = \lambda^{-\frac{1}{2}} X v_1$$

Defiating X,

$$\tilde{X} = X - uu' X = X - \lambda^{\frac{1}{2}} u v_1' = X - X v_1 v_1'$$

$$\tilde{C} = \tilde{X}^T \tilde{X} = \frac{1}{n}(X - uu^T X)^T (X - uu^T X)$$

$$= \frac{1}{n}(X^T X - X^T u u^T X)$$

$$\boxed{= \frac{1}{n}(X^T X - \lambda v_1 v_1^T)}$$

Scanned with CamScanner

D) $Cv_j^o = \lambda_j^o v_j^o$

$\tilde{C} = C - \lambda_j^o v_j^o v_j^{oT}$

$(C - \lambda_j^o v_j^o v_j^{oT}) v_j^o$

$= C v_j^o - \lambda_j^o v_j^o v_j^{oT} v_j^o$

$= \lambda_j^o v_j^o - \lambda_j^o v_j^o (v_j^{oT} v_j^o)$

if $j \neq 1$, then $(C - \lambda_j^o v_j^o v_j^{oT}) v_j^o = \lambda_j^o v_j^o - \lambda_j^o v_j^o (0) = \lambda_j^o v_j^o$

Thus $(C - \lambda_j^o v_j v_j^T) = \tilde{C}$ has the same eigenvectors as C and the same eigenvalues as C except that the largest one has been replaced by 0.

E) Hotelling deflation assumes that the largest eigenvalue λ_1 and an associated eigenvector $v(1)$ of C has been obtained from its defiation matrix $\tilde{C}$, which has the same eigenvalues $\lambda_2 \ldots \lambda_j$ as C except that $\tilde{C}$ has eigenvalue 0 with eigenvector $v(1)$ instead of eigenvalue λ_1.

Scanned with CamScanner

So if u be the first principal eigenvector of $\tilde{C}$

$U = V_2$ because the first eigenvector is zero

(d) After K iterations,

$$C^K u_0 = \lambda_1^K \left[c_1 v_1 + c_2 \left(\frac{\lambda_2}{\lambda_1}\right)^K u_2 + \dots + c_n \left(\frac{\lambda_n}{\lambda_1}\right)^K u_n \right] \quad [u_0 = c_1 v_1 + \dots + c_n v_n]$$

Since here λ_1 dominates, the ratio $\left(\frac{\lambda_i}{\lambda_1}\right)^K \to 0$ as $K \to \infty$ for all i. Therefore, $C^K u_0 = \lambda_1^K c_1 v_1$ gets better as K grows. Here $c_1 \dots c_n$ is constant.

So, for large powers of K we will obtain a good approximation of the dominant eigenvector.

For K iteration, the computational complexity is $O(K^2)$.

In each iteration if we do $u_i = \frac{C u_{i-1}}{\|C u_{i-1}\|_2}$,

this can be a better approach because the

Scanned with CamScanner

power method tends to produce approximations with large entries. In practice it is best to 'scale down' each approximation before proceeding to the next iteration. The way to accomplish this scaling is to determine the component of cv_i that has the largest absolute value.

e) The computational complexity is: kKm^2

(see code)

Scanned with CamScanner

Problem 2

a) We have to set r and c such that $\min \|q - x\| \leq r$ and $\|q - x'\| \leq cr$. More specifically, an r-near query with constant factor c will return a j such that $\|q - x_j'\| \leq cr$ if there is some i so that $\min_i \|q - x_i\| \leq r$.

We can do this by solving nearest neighbor checking r with $r = 1, 2, 4, 8, \ldots R$

We only need to give output if there is a point within r, even if something within cr.

Scanned with CamScanner

b) Given,

$h(x_i) = x_i[a]$

For each point $x_i, x_j \in \{0,1\}^d$:

* $Pr(h(x_i) = h(x_j)) \geq P_1$

where $P_1 = 1 - \frac{r}{d} \approx e^{-r/d}$ if $dist(x_i, x_j) \leq r$.

* $Pr(h(x_i) = h(x_j)) \leq P_2$

where $P_2 = 1 - \frac{cr}{d} \approx e^{-cr/d}$ if $dist(x_i, x_j) \geq cr$

Here $\boxed{P_1 > P_2}$

c) $Pr(g(x_i) = g(x_j)) \geq P_1^K$ [lower bound] $[d(x_i, x_j) \leq r]$

$Pr(g(x_i) = g(x_j)) \leq P_2^K$ [upper bound] $[d(x_i, x_j) \geq cr]$

d) $Pr(g_b(x_i) = g_b(x_j)) \geq 1 - \frac{1}{n}$ [LB] $[d(x_i, x_j) \leq r]$

$Pr(g_b(x_i) = g_b(x_j)) \leq \frac{1}{n}$ [UB] $[d(x_i, x_j) \geq cr]$

Scanned with CamScanner

e) $K = \frac{\ln(n)}{\ln(1/P_2)}$

This implies that for fixed b, the expected numbers of x' that map to the same bucket as q is at most $n \times \frac{1}{n} = 1$.

Applying linearity of expectation over all l buckets, the expected number of false positive is at most l.

By Markov's inequality the probability there are more than $4l$ false positives is therefore at most $\frac{1}{4}$ and at least $\boxed{\frac{3}{4}}$.

According to first event, for any b

$$Pr[g_b(x^*) \neq g_b(q)] \leq 1 - P_1^K = 1n^{-\rho} = 1 - \frac{1}{n^\rho} \quad [K = \frac{\ln(n)}{\ln(1/P_2)}$$

$$\rho = \frac{\ln(P_1)}{\ln(P_2)}$$

As $l = n^\rho$,

$$Pr[g_b(x^*) \neq g_b(q) \forall i] \leq (1 - \frac{1}{n^\rho})^{n^\rho} \leq \frac{1}{e}$$

So, the first event holds with probability at least $\boxed{1 - \frac{1}{e}}$

Scanned with CamScanner

with union bound the probability of both event holding at least: $\boxed{1-\frac{1}{4}-1/e \geq \frac{1}{3}}$

2f) In expectation, there are $O(L)$ points in the data set which map to same hash functions as q for some K. we need to examine these points to check if any of them within distance cr from q

No, it is not guranteed that there is a point with distance at most cr because the probability of upper bound is very low (only 1/n).

Scanned with CamScanner