Linear Algebra

Yoav Shoshan

① A symetric. $\forall v \; v^TAv \geq 0$.
spectral decomposition: $A = QDQ$
Q contain all of A's eigenvectors
D diagonal and contain corresponding eigenvalues, λ_i.

$$\Longrightarrow \quad 0 \leq v^TAv = v^TQD(\lambda)Q^Tv = \sum_i \lambda_i |Q^Tv|_i^2$$

$$\Longrightarrow \quad \lambda_i \geq 0 \quad \text{for all } i \quad (\text{if previous true for } \forall v)$$

$$\Longrightarrow \quad A = QD(\lambda)Q^T = QD(\sqrt{\lambda})D(\sqrt{\lambda})Q^T$$

$$\equiv XX^T \quad (X = QD(\sqrt{\lambda})) \quad //$$

Conversly, if $A = XX^T$, Then

$$v^TAv = v^TXX^Tv \equiv w^Tw = \langle w, w \rangle \geq 0 \quad //$$

② A, B are PSD.
define $C = \theta A + (1-\theta)B$ for $\theta \in [0,1]$
Then; for $\forall v$:

$$v^TCv = v^T(\theta A + (1-\theta)B)v =$$

$$\underbrace{\theta v^TAv}_{\geq 0} + \underbrace{(1-\theta)}_{\geq 0}\underbrace{v^TBv}_{\geq 0} \geq 0 \Longrightarrow C \text{ is PSD} \quad //$$

Calculus & Probability

① $x^TAx = x_iA_{ij}x_j$ (Einstien notation of summation)

$$\frac{\partial}{\partial x_k}(x_iA_{ij}x_j) = \frac{\partial x_i}{\partial x_k}A_{ij}x_j + x_i\frac{\partial A_{ij}x_j}{\partial x_k}$$

$$= \delta_{ik}A_{ij}x_j + x_iA_{ij}\frac{\partial x_j}{\partial x_k} =$$

$$= \delta_{ik}A_{ij}x_j + x_iA_{ij}\delta_{jk} = A_{kj}x_j + x_iA_{ik}$$

$$= A_{ki}x_i + A_{ik}x_i = (Ax + A^Tx)_k \quad //$$

② $p = (p_1 -- p_n)$ $\quad \sum_i p_i = 1 \quad p_i \geq 0$

$$H(p) = -\sum_i p_i \log p_i$$

$$G(p,\lambda) = H(p) - \lambda\sum_i p_i$$

(We would like to find p_i That minimize the entropy given the constraint

I $\frac{\partial G}{\partial p_i} = -1 - \ln p_i - \lambda \overset{!}{=} 0$

II $\frac{\partial G}{\partial \lambda} = -\sum_i p_i = 1$

I: $\ln p_i = -1 - \lambda \Rightarrow p_i = e^{-1-\lambda}$

II: $1 = -\sum_i e^{-1-\lambda} = Ne^{-1-\lambda} \Rightarrow$

$e^{-1-\lambda} = \frac{1}{N} \Rightarrow \lambda = -1 - \ln\frac{1}{N}$

I: $-1 - \ln p_i + 1 + \ln\frac{1}{N} = 0$

$\Rightarrow p_i = \frac{1}{N}$ // uniform dist.

(3) (a) $P[X_0 \geq \max(X_1, \dots X_{n-1})] =$ Total prob.

$X_i \in [0,\infty)$ $= \int_0^\infty P(X_0 \geq \max(X_1 \dots X_{n-1}) | X_0 = a) \, f_{X_0}(a) \, da$

$$\int_0^\infty P(a \geq \max(X_1 \dots X_{n-1})) \, f_{X_0}(a) \, da =$$

$$\int_0^\infty P\left(\bigcap_{i=1}^{n-1} X_i \leq a\right) f_X(a) \, da \underset{\uparrow}{=} \int \prod_{i=1}^{n-1} P(X_i \leq a) \, f_{X_0}(a) \, da$$

independent

$$\underset{\text{identical}}{=} \int_0^\infty (P(X_0 \leq a))^{n-1} f_X(a) \, da = \int_0^\infty (F_{X_0}(a))^{n-1} f_{X_0}(a) \, da$$

(b) $F_{X_0}(a) = \int_0^a f_X(a') \, da'$, $f_X(a) = \frac{\partial F_X}{\partial x}\Big|_{x=a} \equiv F'_X(a)$

$$\int_0^\infty (F_{X_0}(a))^{n-1} F'_X(a) \, da = \int_0^\infty \frac{1}{n}\left[(F_X(a))^n\right]' \, da$$

$$= \frac{1}{n} F_X^n(a)\Big|_0^\infty = \frac{1}{n}(1-0) = \frac{1}{n}$$

$$\Longrightarrow P(X_0 \geq \max(X_1 \dots X_{n-1})) = \frac{1}{n}$$

Decision Rules & Concentration Bounds

(1) $L(h) = \sum_{x,y} P(X=x, Y=y)\, \Delta_{zo}(y, h(x))$

$\sum_{x} P(X=x) \left(\sum_{y=0}^{L} P(Y=y \mid X=x)\, \Delta_{zo}(y, h(x)) \right)$

$\Rightarrow$ for a given x

$L(h) \propto \sum_{y=0}^{L} P(Y=y \mid X=x)\, \Delta_{zo}(y, h(x)) \overset{h(x)=y'}{=}$

$\sum_{y=0}^{L} P(Y=y \mid X=\hat{x}) - P(Y=y' \mid X=x) \qquad h(x)=y'$

$\Rightarrow$ minimizing $L(h) \Leftrightarrow$ maximizing $P(Y=y' \mid X=x)$

$\Rightarrow$ we find y for wich $P(Y=y \mid X=x)$ is the largest

$\Rightarrow h(x) = \arg\max_{y \in [0, \ldots L]} P(Y=y \mid X=x)$

(2) $f(\vec{x},\vec{\mu},\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$

(a) $P(y=1|x) > P(y=0|x) \implies y=1$

Bayes: $\frac{f_x(x|Y=1)P(Y=1)}{f_x(x)} > \frac{f_x(x|Y=0)P(Y=0)}{f_x(x)}$

$$\implies f_x(x|Y=1)P(Y=1) > f_x(x|Y=0)P(Y=0)$$

$$p\cdot\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right) > (1-p)\exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)$$

$$\frac{p}{1-p} > \exp\left(-\frac{1}{2}\left[(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) - (x-\mu_1)^T\Sigma^{-1}(x-\mu_1)^T\right]\right)$$

$$= \exp\left(-\frac{1}{2}\left[x^T\Sigma^{-1}(\mu_1-\mu_0) + (\mu_1-\mu_0)^T\Sigma^{-1}x + \mu_0^T\Sigma^{-1}\mu_0 - \mu_1\Sigma^{-1}\mu_1\right]\right)$$

$$2\ln\left(\frac{1-p}{p}\right) < x^T\Sigma^{-1}(\mu_1-\mu_0) + (\mu_1-\mu_0)^T\Sigma^{-1}x + \mu_0^T\Sigma^{-1}\mu_0 - \mu_1\Sigma^{-1}\mu_1$$

(for sym A.

$A^{-1}A = I \implies (A^{-1}A)^T = A^T(A^{-1})^T = A(A^{-1})^T = I$

$\implies A^{-1} = (A^{-1})^T$, A^{-1} symmetric)

Σ is cov matrix, thus sym, $\implies \Sigma^{-1}$ also sym

$$\implies x^T\Sigma^{-1}(\mu_1-\mu_0) = (\mu_1-\mu_0)^T\Sigma^{-1}x$$

$$\implies \boxed{(\mu_1-\mu_0)^T\Sigma^{-1}x > \ln\left(\frac{1-p}{p}\right) + \frac{\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0}{2}}$$

(2)(3) for d Dimensions the decision boundary should be a d-1 surface on which $P(y=1|X) = P(y=0|X)$, we got a d-1 degrees of freedom in the equation in section a

(3) $S = \sum_{i=1}^{n} X_i$, $X_i \sim U(-3,5)$

$S = n\bar{X}$

$P(S > n^2 + 0.2n) = P(\bar{X} > n + 0.2) =$ (Uniform dist $\mu = \frac{5-3}{2} = 1$)

$= P(\bar{X} - \mu > n - 0.8) \le P(|\bar{X} - \mu| \ge n - 0.8)$

$\le 2\exp\left(\frac{-2n(n-0.8)^2}{8^2}\right) \overset{!}{<} 0.1$

$\Longrightarrow n(n-0.8)^2 > -32\ln(0.05)$

$n(n-0.8)^2 + 32\ln(0.05) > 0$

real root by wolfram is 5.12

so we choose $n \ge 6$ //

(4) (a) $E(R_{ij}) = \frac{1}{m}L_j + \frac{m-1}{m}\cdot 0 = \frac{1}{m}L_j$ (uniform dist. for servers ass.)

$E(R_i) = E\left(\sum_j R_{ij}\right) = \sum_j E(R_{ij}) = \sum_j \frac{1}{m}L_j = \frac{L}{m}$

(b) $P(R_i \ge (1+\delta)E(R_i)) = \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\frac{L}{m}} \equiv C$

(c) $P\left(\bigcup_{i=1}^{m} R_i \ge (1+\delta)E(R_i)\right) = P\left(\bigcup_{i=1}^{m} R_i \ge (1+\delta)\frac{L}{m}\right)$

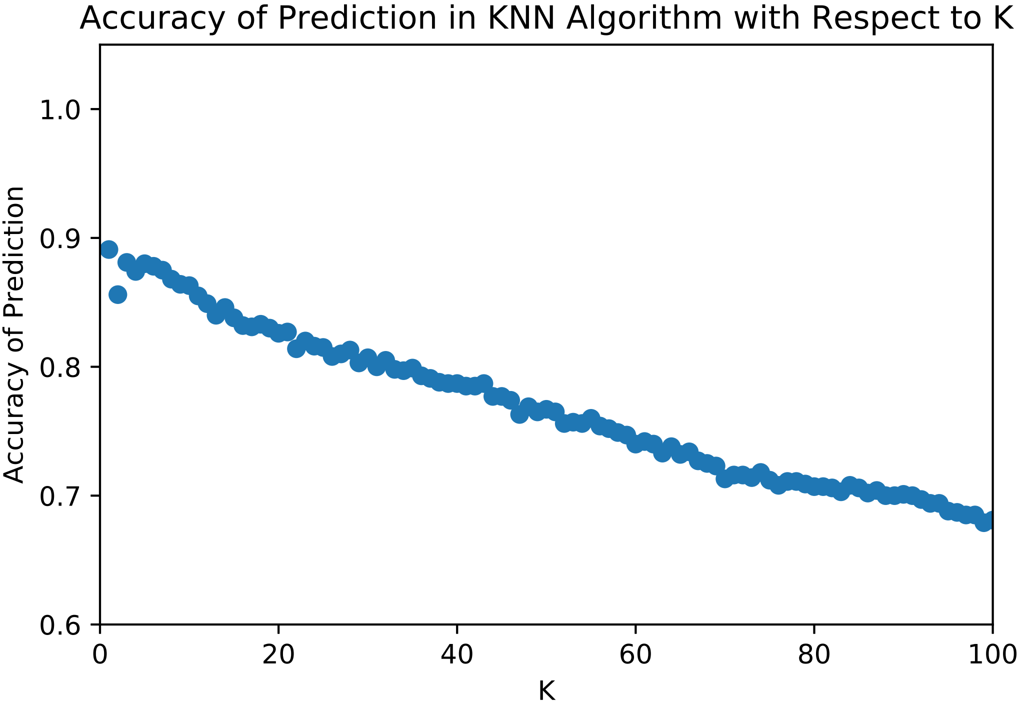
$= \sum_i P\left(R_i \ge (1+\delta)\frac{L}{m}\right) = \sum_i C = mC$ //

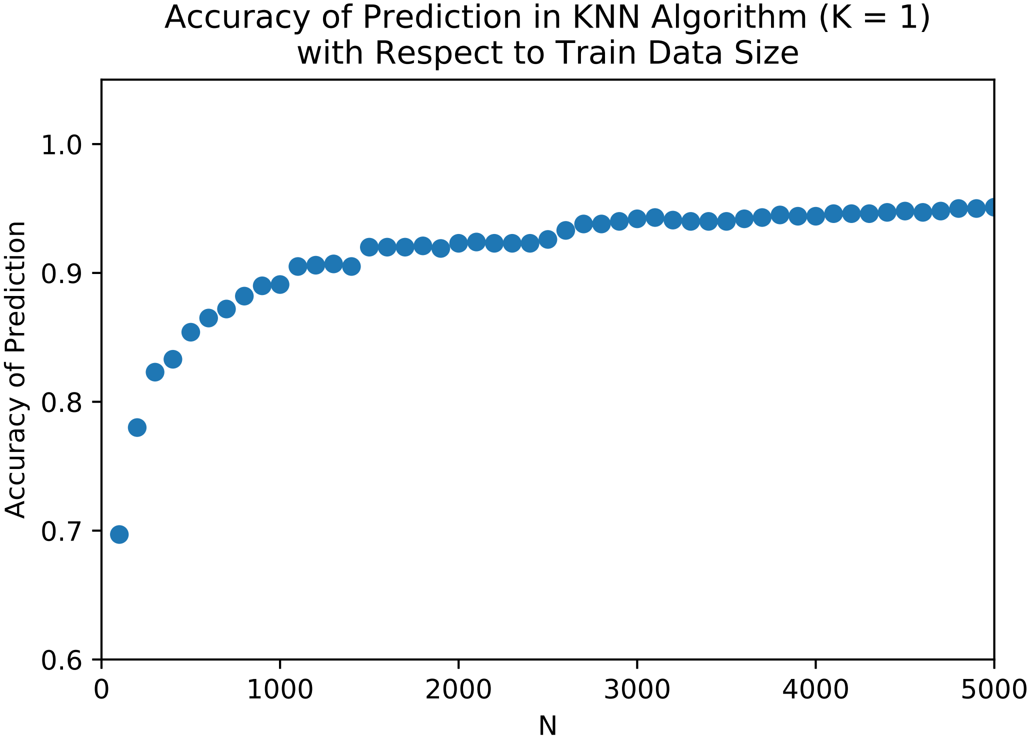
Programming Assignment (Discussion)

(b) Running knn for K=10 and a train data of size 1000 we got accuracy of ~ 0.87. meaning 87% of all test images were classified correctly. The expected accuracy of a random predictor should be ~0.1, there are 10 possible labels and each label has an equal prob to be chosen.

(c) we got the best acurracy for K=1, and it is easy to see that the accuracy decreases with K. it is possible that for this dataset, the nearest neighbor for most images will carry the same label, but still, the images can't be perfectly clustered in an euclidian space so we are likely to fined many images with different labeles in the neighborhood of a given image.

d) we can see that the accuracy increases with N. meaning the more train data we have, the more likely we are to include relativly close neighbors with labels identical to that of a given image.


Accuracy of Prediction in KNN Algorithm with Respect to K
Accuracy of Prediction
1.0
0.9
0.8
0.7
0.6
0
20
40
60
80
100
K



Accuracy of Prediction in KNN Algorithm (K = 1)
with Respect to Train Data Size
Accuracy of Prediction
1.0
0.9
0.8
0.7
0.6
0
1000
2000
3000
4000
5000
N