آرین سنکر - تکلیف شماره ۴ درس شناسایی آماری الگو

6.3. A vector X in the original N-dimensional space is described as: $X = \sum_{i=0}^{N-1} y(i) e_i$, $y(i) = e_i^T X$ (eq 6.14)

Moreover, the transformation into the $M \leq N$ dimensional subspace that we assume W.L.O.G., spans the first M basis vectors $e_0 \ldots e_{M-1}$ is defined by: $\hat{X} = \sum_{i=0}^{M-1} y(i) e_i$

$$\rightarrow X - \hat{X} = \sum_{i=M}^{N-1} y(i) e_i$$

The expectation of the squared error gives us the MSE:

$$E[\| X - \hat{X} \|^2] = E\left[\sum_{i=M}^{N-1} \sum_{j=M}^{N-1} y(i)^* y(j) e_i^T e_j \right] \quad (I)$$

However, from the orthogonality of the basis functions we know that $e_i^T e_j = \delta_{ij}$ where δ_{ij} is the Kronecker's Delta.

therefore we can rewrite eq. (I) as : $E[\| X - \hat{X} \|^2] = \sum_{i=M}^{N-1} E[y(i)^2]$

$$= \sum_{i=M}^{N-1} E[e_i^T x x^T e_i] \xrightarrow[\text{are constants w.r.t. expectation}]{e_i \text{ and } e_i^T} \sum_{i=M}^{N-1} e_i^T E[x x^T] e_i = \sum_{i=M}^{N-1} e_i^T R_x e_i$$

defined as R_X the auto correlation matrix

Therefore the objective is to minimize $\sum_{i=M}^{N-1} e_i^T R_x e_i$. If we set all e_is to 0 this term is obviously minimized. Thus, in order to avoid this trivial solution we introduce the constraint $e_i^T e_i = 1$ as hinted by the problem statement.

In order to solve the new constrained minimization task we construct the Lagrangian:

$$\mathcal{L} = \sum_{i=M}^{N-1} e_i^T R_x e_i - \sum_{i=0}^{N-1} \lambda_i (e_i^T e_i - 1)$$

a) if we set the derivative of the Lagrangian w.r.t. e_i , $i \in [M, N-1]$ to zero we get: $\frac{\partial \mathcal{L}}{\partial e_i} = 2 R_x e_i - 2 \lambda_i e_i = 0 \rightarrow R_x e_i = \lambda_i e_i$

e_i's are the eigenvectors of R_X and λ_i's are the eigenvalues of R_X

(b) if we substitute $R_x e_i = \lambda_i e_i$ in the original formula derived for MSE

we get: $E[\|x-\hat{x}\|^2] = \sum_{i=M}^{N-1} e_i^T R_x e_i = \sum_{i=M}^{N-1} \lambda_i e_i^T e_i$ The constraint $e_i^T e_i$

implies: $E[\|x-\hat{x}\|^2] = \sum_{i=M}^{N-1} \lambda_i$ (II)

Moreover we know that R_x is positive semidefinite because:

$$\forall y \neq 0, \quad y^T E(xx^T) y = \frac{1}{d} y^T x x^T y = \frac{1}{d} \underbrace{(y^T x)}_{A} \underbrace{(y^T x)^T}_{A}$$

$$= \frac{1}{d} AA^T \geq 0$$

$\Rightarrow$ eigenvalues of R_x are non-negative $\longrightarrow$ This means for eq (II) to be minimized we must take smallest eigenvalues $\lambda_m, \ldots, \lambda_{N-1}$ from R_x.

Consecutively this means that the original M-dimensional subspace must consist of the M largest eigenvalues of R_x, $\lambda_1 \ldots \lambda_M$ (and be the subspace spanned by their respective eigenvectors $e_1, \ldots, e_M$).

moreover, the sum of the variances of the componen'ts is equal to:

$$\sum_{i=0}^{M-1} E[(x_i-\mu_i)^2] \quad \text{(III)}$$

W.L.O.G. we can assume that we have subtracted the mean of each component from it so that all x_i's are zero-centered. with that we can rewrite (III) as:

$$\sum_{i=0}^{M-1} E[x_i^2] = \sum_{i=0}^{M-1} e_i^T E[xx^T] e_i$$

(this is because of the same logic we follow in the first part of the problem)

and we have shown that $\sum_{i=0}^{M-1} e_i^T E[xx^T] e_i = \sum_{i=0}^{M-1} \lambda_i$

following from the consequence of the previous part, the choice of λ_i's maximizes $\sum_{i=0}^{M-1} \lambda_i$ (and minimizes $\sum_{i=M}^{N-1} \lambda_i$).

6.4.

a) $\hat{x} = \sum_{i=0}^{M-1} y(i) e_i + \sum_{i=M}^{N-1} c_i e_i \quad , \quad x = \sum_{i=0}^{N-1} y(i) e_i \quad , \quad y(i) = e_i^T x$

$$\rightarrow E\left[\| x - \hat{x} \|^2\right] = E\left[\| \sum_{i=0}^{N-1} y(i) e_i - \sum_{i=0}^{M-1} y(i) e_i - \sum_{i=M}^{N-1} c_i e_i \|^2\right]$$

$$= E\left[\| \sum_{i=M}^{N-1} (y(i) - c_i) e_i \|^2\right] = E\left[\sum_{i=M}^{N-1} \sum_{j=M}^{N-1} (y(i) - c_i)(y(j) - c_j) e_i^T e_j\right] \quad (I)$$

but we know that $e_i^T e_j = \delta_{ij} \xrightarrow{(I)} = E\left[\sum_{i=M}^{N-1} (y(i) - c_i)^2\right]$

$= \sum_{i=M}^{N-1} \left(E[y(i)^2] - 2E[y(i)]E[c_i] + E[c_i^2]\right) \rightarrow c_i$ is a constant therefore

$\rightarrow = \sum_{i=M}^{N-1} \left(E[y(i)^2] - 2E[y(i)] c_i + c_i^2 \right)$

in order to minimize this w.r.t. choice of c_i, we set the derivative w.r.t c_i equal to zero:

$$\frac{\partial}{\partial c_i} \sum_{i=M}^{N-1} \left(E[y(i)^2] - 2E[y(i)] c_i + c_i^2 \right) = -2E[y(i)] + 2c_i = 0$$

$$\Rightarrow c_i = E[y(i)], \; M \leqslant i \leqslant N-1$$

b) Now that we have established the best values for c_i's we can try to decide for the direction of e_i's. we showed in the previous section that:

$$E[\| x - \hat{x} \|^2] = E\left[\sum_{i=M}^{N-1} (y(i) - c_i)^2\right] \quad (II)$$

using the results from the previous section:

$$(II) = E\left[\sum_{i=M}^{N-1} (y(i) - E[y(i)])^2\right] = E\left[\sum_{i=M}^{N-1} (e_i^T x - e_i^T E[x])^2\right]$$

$$= E\left[\sum_{i=M}^{N-1} (e_i^T (x - E[x]))^2\right] = E\left[\sum_{i=M}^{N-1} e_i^T (x - E[x])(x - E[x])^T e_i\right]$$

$$= \sum_{i=M}^{N-1} e_i^T E\left[(x - E[x])(x - E[x])^T\right] e_i = \sum_{i=M}^{N-1} e_i^T \Sigma_x e_i$$

where we have used the definition $\underline{E\left[(x - E[x])(x - E[x])^T\right] = \Sigma_x}$

in order to minimize this w.r.t. e_is we need to introduce a constraint of orthonormality. I.e. that $e_i^T e_i = 1$ AND $e_i^T e_j = 0$, $i \neq j$. we construct the Lagrangian:

$$L = \sum_{i=M}^{N-1} e_i^T \Sigma_x e_i - \sum_{i=0}^{N-1} \lambda_i (e_i^T e_i - 1)$$

$$\rightarrow \frac{\partial L}{\partial e_i} = 2\Sigma_x e_i - 2\lambda_i e_i \rightarrow \Sigma_x e_i = \lambda_i e_i \rightarrow \begin{cases} e_i := \text{eigenvectors of } \Sigma_x \\ \lambda_i := \text{eigenvalues of } \Sigma_x \end{cases}$$

C) Substituting $\Sigma_x e_i = \lambda_i e_i$ in $\sum_{i=M}^{N-1} e_i^T \Sigma_x e_i$ we get:

$$E[\|x - \hat{x}\|^2] = \sum_{i=M}^{N-1} e_i^T \Sigma_x e_i = \sum_{i=M}^{N-1} \lambda_i e_i^T e_i \xrightarrow{e_i^T e_i = 1} \boxed{\sum_{i=M}^{N-1} \lambda_i}$$

$\rightarrow$ in order to minimize $E[\|x - \hat{x}\|^2]$ we take $(N-1-M+1 = N-M)$ smallest eigenvalues of Σ_x.

9.1. ———

9.2. $P_i(d) = (p(i|i))^{d-1}(1 - p(i|i))$

average duration for staying in state i:

$$\bar{d} = \sum_{d=0}^{\infty} P_i(d) \times d = \sum_{d=0}^{\infty} d\,(p(i|i))^{d-1} (1 - p(i|i))$$

$$= (1 - p(i|i)) \sum_{d=0}^{\infty} d\,(p(i|i))^{d-1} \qquad (I)$$

we know from geometric series sum that:

$$1 + x + x^2 + \dots + x^n = \frac{1 - x^{n+1}}{1 - x} \quad \text{and when } |x| < 1 \text{ and } n \to +\infty$$

we get: $\sum_{i=0}^{\infty} x^i = \frac{1}{1-x} \xrightarrow[\text{from both sides}]{\text{derivative}} \sum_{i=0}^{\infty} i x^{(i-1)} = \frac{1}{(1-x)^2}$

setting $p(i|i) = x$, eq (I) becomes:

$$\bar{d} = (1 - p(i|i)) \left(\frac{1}{1 - p(i|i)}\right)^2 = \frac{1}{1 - p(i|i)}$$

which is valid because $|p(i|i)| < 1$

9.3. According to the reestimation formulae derived within the textbook we know that each estimation has an intuitive representation:

$$\bar{P}(j|i) = \frac{E[\#\text{ of transitions from } i \text{ to } j]}{E[\#\text{ of transitions from } i]} \quad (\mathrm{I})$$

$$\bar{P}_x(r|i) = \frac{E[\#\text{ at state } i \text{ and emits } r]}{E[\#\text{ of beig at state } i]} \quad (\mathrm{II})$$

this means that if we have Q different versions of the same pattern, the reestimation formulae stay the same as before except that the mathematical interpretation of their intuitive representations will now differ (eq I and II).

That is, in order to find any of the expectations above we now also have to sum over Q.

For example let's evaluate $E[\#\text{ of transitions from } i \text{ to } j]$:

originally this would equal $\sum_{k=1}^{N-1} \xi_k(i,j)$ when only one pattern was present. However we now have Q different versions of that same pattern.

If we want to derive the same formula, it is obvious that:

$$\begin{aligned}E[\#\text{ of transitions from } i \text{ to } j] = {} & E[\#\text{ of transitions from } i \text{ to } j \text{ in version 1 of pattern}] \\ & + E[\#\text{ of transitions from } i \text{ to } j \text{ in version 2 of pattern}] \\ & + \cdots \\ & + E[\#\text{ of transitions from } i \text{ to } j \text{ in version Q of pattern}]\end{aligned}$$

$$\rightarrow E[\#\text{ of transitions from } i \text{ to } j] = \sum_{i=1}^{Q}\sum_{k=1}^{N_m-1} \xi_k(i,j) \quad \text{or} \quad \sum_{i=1}^{Q} \frac{1}{P(X_m|S)} \sum_{k=1}^{N_m-1} \xi_k(i,j,X_m|S)$$

using the definition of $\xi_k(i,j)$

all of the other expectations can be calculated following the exact same logic.

9.4.

$$\bar{P}(j|i) = \frac{\sum_{k=1}^{N-1} \xi_k(i,j)}{\sum_{k=1}^{N-1} \gamma_k(i)}$$ → using definitions for $\xi_k(i,j)$ and $\gamma_k(i)$

$$\rightarrow \bar{P}(j|i) = \frac{\sum_{k=1}^{N-1} \alpha(i_k=i) P(j|i) P(x_{k+1}|j) \beta(i_{k+1}=j)}{\sum_{k=1}^{N-1} \alpha(i_k=i)\beta(i_k=i)}$$

Because we need a $\beta(i_{k+1}=i)$ in the denominator we use the recursion relation for β:

$$\beta(i_k) = \sum_{i_{k+1}} \beta(i_{k+1}) P(i_{k+1}|i_k) P(\underline{x}_{k+1}|i_{k+1})$$

$$\rightarrow \bar{P}(j|i) = \frac{\sum_{k=1}^{N-1} \alpha(i_k=i) P(j|i) P(x_{k+1}|j) \beta(i_{k+1}=j)}{\sum_{k=1}^{N-1} \alpha(i_k=i) \sum_{j=1}^{N} P(j|i) P(\underline{x}_{k+1}|j)\beta(i_{k+1}=j)}$$

Now we can replace α and β with their scaled versions. If we do so the c_k and $\frac{1}{c_k}$ cancel each other out, amounting to $\underline{1}$. in the numerator and in the denominator

Therefore the formula stays unchanged.

Similarly for $\bar{P}_{\underline{x}}(r|i)$ we will arrive at two formulae for the numerator and the denominator

where c_k and $\frac{1}{c_k}$ cancel out and we get the same result as before.

→ the formulae don't change with this scaling scheme.