Question | सफेदी की चमकार ... ज्यादा सफेद

The Whitening Transform

This is given by $\boxed{\tilde{r}_i = \Lambda^{-1/2} U^T \underline{p}_i}$. Here, we deal with

$k \times 1$ Type I normalised pattern vectors (i.e., normalised with respect to the mean) $\underline{p}_i$, $0 \le i \le n-1$. Λ is a diagonal matrix of their eigenvalues, and U is their corresponding eigenvector matrix. What is the covariance matrix of the transformed patterns $\tilde{r}_i$? (The following will not be evaluated: Why do you think this transformation is called so?)

$$\tilde{r}_i = \Lambda^{-1/2} U^T \underline{p}_i$$

Stacking all n of them together

$$\tilde{R} = \Lambda^{-1/2} U^T P$$

$$\begin{aligned}
\text{Cov} \triangleq \frac{1}{n} \tilde{R}\tilde{R}^T &= \frac{1}{n} \Lambda^{-1/2} U^T P (\Lambda^{-1/2} U^T P)^T \\
&= \frac{1}{n} \Lambda^{-1/2} U^T (P P^T) U \Lambda^{-1/2} \\
& \qquad (\because U^{T^T} = U,\ \Lambda^{-1/2^T} = \Lambda^{-1/2} : \text{diag}) \\
&= \Lambda^{-1/2} \underbrace{U^T A U}_{U^{-1}AU = \Lambda} \Lambda^{-1/2} \\
&= \Lambda^{-1/2} \Lambda \Lambda^{1/2} \\
&= I \qquad \text{(Multiplication of diagonal matrices = multiplication of individual terms)}
\end{aligned}$$

Why is this called so? No correlation, same variance along each direction ~ white light / white noise.

Question: Orthopaedic Orthonormality: Bone-Breaking Normal work

Show that for non-repeated eigenvalues of a symmetric matrix, the eigenvectors will be ortho-normal. Consider the case of repeated eigenvalues of a symmetric matrix A. Specifically use the Gram Schmidt orthogonalisation process to show that it is still possible to obtain a set of orthonormal vectors.

Let the symmetric matrix be $A_{n \times n}$ with n eigenvalues λ_k and corresponding eigenvectors $\underline{u}_k$ i.e., $A\underline{u}_k = \lambda_k \underline{u}_k$

Further, given an orthogonal set, it is always possible to convert it to an orthonormal one by dividing the vector by its norm e.g., Euclidean.

Non-repeated eigenvalues

Consider $\lambda_i \; \underline{u}_i \cdot \underline{u}_j = \lambda_i \; \underline{u}_i^T \underline{u}_j$

$= (\lambda_i \underline{u}_i)^T \underline{u}_j$ (transpose of a scalar)

$= (A\underline{u}_i)^T \underline{u}_j$

$= \underline{u}_i^T A^T \underline{u}_j = \underline{u}_i^T A \underline{u}_j$ (transpose of a symmetric matrix)

$= (\underline{u}_i^T)(A\underline{u}_j) = \underline{u}_i^T \lambda_j \underline{u}_j$

$= \lambda_j \; \underline{u}_i^T \underline{u}_j$

$\Rightarrow \lambda_i \; \underline{u}_i^T \underline{u}_j = \lambda_j \; \underline{u}_i^T \underline{u}_j \Rightarrow \underbrace{(\lambda_i - \lambda_j)}_{\neq 0 \text{ as non-repeated eigenvalues}} \; \underbrace{\underline{u}_i^T \underline{u}_j = 0}_{\Rightarrow \underline{u}_i \,\&\, \underline{u}_j \text{ are orthogonal.}}$

Repeated eigenvalues of degree k (say)

FIRST, [consider the Gram-Schmidt orthogonalisation construction

Given a set of k basis vectors $\underline{v}_1, \underline{v}_2, \dots \underline{v}_k$

(Basis vectors $\Rightarrow$ they are linearly independent)

To construct an orthogonal set $\underline{u}_1, \underline{u}_2, \dots \underline{u}_k$ from it.

Step 0: Take $\underline{u}_1 = \underline{v}_1$

Step I: To construct $\underline{u}_2 \perp \underline{u}_1$ such that the span of $\underline{u}_1, \underline{u}_2$ = span of $\underline{v}_1, \underline{v}_2$

Take $\underline{u}_2 = a_1 \underline{u}_1 + \underline{v}_2$ (linear combination)

To find a_1: take a dot product with $\underline{u}_1$

$$\underbrace{\underline{u}_2 \cdot \underline{u}_1}_{=0} = a_1 (\underline{u}_1 \cdot \underline{u}_1) + \underline{v}_2 \cdot \underline{u}_1$$

$$\Rightarrow a_1 = -\frac{\underline{v}_2 \cdot \underline{u}_1}{\underline{u}_1 \cdot \underline{u}_1} = -\frac{\langle \underline{v}_2, \underline{u}_1 \rangle}{\langle \underline{u}_1, \underline{u}_1 \rangle}$$

$$\Rightarrow \underline{u}_2 = \underline{v}_2 - \frac{\langle \underline{v}_2, \underline{u}_1 \rangle}{\langle \underline{u}_1, \underline{u}_1 \rangle}$$

Step II: To construct $\underline{u}_3 \perp (\underline{u}_1, \underline{u}_2)$ such that the span of $\underline{u}_1, \underline{u}_2, \underline{u}_3$ = span of $\underline{v}_1, \underline{v}_2, \underline{v}_3$

Take $\underline{u}_3 = a_1 \underline{u}_1 + a_2 \underline{u}_2 + \underline{v}_3$

To find a_1, a_2: take dot products with $\underline{u}_1, \underline{u}_2$

$$\rightarrow \underbrace{\underline{u}_1 \cdot \underline{u}_3}_{=0} = a_1 \underline{u}_1 \cdot \underline{u}_1 + a_2 \underbrace{\underline{u}_2 \cdot \underline{u}_1}_{=0} + \underline{v}_3 \cdot \underline{u}_1$$

$$\Rightarrow a_1 = -\frac{\underline{v}_3 \cdot \underline{u}_1}{\underline{u}_1 \cdot \underline{u}_1} = -\frac{\langle \underline{v}_3, \underline{u}_1 \rangle}{\langle \underline{u}_1, \underline{u}_1 \rangle}$$

$$\rightarrow \underbrace{\underline{u}_2 \cdot \underline{u}_3}_{=0} = a_1 \underbrace{\underline{u}_1 \cdot \underline{u}_2}_{=0} + a_2 \underline{u}_2 \cdot \underline{u}_2 + \underline{v}_3 \cdot \underline{u}_2$$

$$\Rightarrow a_2 = -\frac{\underline{v}_3 \cdot \underline{u}_2}{\underline{u}_2 \cdot \underline{u}_2} = -\frac{\langle \underline{v}_3, \underline{u}_2 \rangle}{\langle \underline{u}_2, \underline{u}_2 \rangle}$$

$$\Rightarrow u_3 = v_3 - \frac{\langle v_3, u_1\rangle}{\langle u_1, u_1\rangle} u_1 - \frac{\langle v_3, u_2\rangle}{\langle u_2, u_2\rangle} u_2$$

Step (k):
$$\boxed{u_k = v_k - \sum_{j=1}^{k} \frac{\langle v_k, u_j\rangle}{\langle u_j, u_j\rangle} u_j}$$

THE MOST IMPORTANT STEP : To show that $u_1, u_2, \ldots u_k$ are also eigenvectors of A, like $v_1, v_2, \ldots v_k$

Step (0) : $u_1 = v_1 \Rightarrow u_1$ is an eigenvector of A

Step (I) $u_2 = v_2 - \frac{\langle v_2, u_1\rangle}{\langle u_1, u_1\rangle} u_1$

$$\Rightarrow A u_2 = A v_2 - \frac{\langle v_2, u_1\rangle}{\langle u_1, u_1\rangle} A u_1$$

$$= \lambda_2 v_2 - \frac{\langle v_2, u_1\rangle}{\langle u_1, u_1\rangle} \lambda_1 u_1$$

Now, $\lambda_1 = \lambda_2 = \cdots \lambda_k$ for all, $= \lambda$ (say)

$$= \lambda \left[v_2 - \frac{\langle v_2, u_1\rangle}{\langle u_1, u_1\rangle} u_1 \right]$$

$\Rightarrow A u_2 = \lambda u_2 \Rightarrow u_2$ is also an eigenvector of A

Step (II) $u_3 = v_3 - \frac{\langle v_3, u_2\rangle}{\langle u_2, u_2\rangle} u_2 - \frac{\langle v_3, u_1\rangle}{\langle u_1, u_1\rangle} u_1$

$$\Rightarrow A u_3 = A v_3 - \frac{\langle v_3, u_2\rangle}{\langle u_2, u_2\rangle} A u_2 - \frac{\langle v_3, u_1\rangle}{\langle u_1, u_1\rangle} A u_1$$

$$= \lambda_3 v_3 - \frac{\langle v_3, u_2\rangle}{\langle u_2, u_2\rangle} \lambda_2 u_2 - \frac{\langle v_3, u_1\rangle}{\langle u_1, u_1\rangle} \lambda_1 u_1$$

Now, $\lambda_1 = \lambda_2 = \cdots \lambda_k$ for all, $= \lambda$ (say)

$$= \lambda \left[v_3 - \frac{\langle v_3, u_2\rangle}{\langle u_2, u_2\rangle} u_2 - \frac{\langle v_3, u_1\rangle}{\langle u_1, u_1\rangle} u_1 \right]$$

$= \lambda u_3 \Rightarrow u_3$ is also an eigenvector of A

Step (k) follows

Question | Projection Intersection: Nothing to lose?

What is the physical significance of projecting a new vector onto an eigenspace? Explain using mathematical expressions, what the above implies, for both the KL-Transform, as well as the SVD. Now, explain how the KL-Transform and the SVD respectively perform lossy compression.

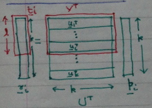
Projecting onto an eigenspace:

KL-Transform: Taking a dot product with the eigenvectors of the cov matrix $\frac{1}{n}PP^T$

SVD: Taking a dot product with the orthonormal basis vectors $\underline{u}_i$, NOT the eigenvectors of P^TP (see the figures below, for more detail on the notation)

Lossy Compression

KL-Transform



$\underline{r}_i = U^T \underline{p}_i$

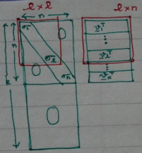
$\underline{t}_i = V^T \underline{p}_i$

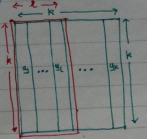
Choosing l:

$$\frac{\sum_i^{l} \lambda_i}{\sum_i^{k} \lambda_i} \geq 0.95$$

min l such that this is valid

SVD





Choosing l:

$$\frac{\sum_i^{l} \sigma_i}{\sum_i^{k} \sigma_i} \geq 0.95$$

min l such that this is valid.

2. Likelihood:

$$p(\underline{t}|X, \underline{w}) = \prod_{n=1}^{N} y_n^{t_n} (1-y_n)^{1-t_n}$$

where $y_n = y(\underline{x}_n) = \sigma(\underline{w}^T \underline{x}_n)$

If the data set is separable, any decision boundary will have the property:

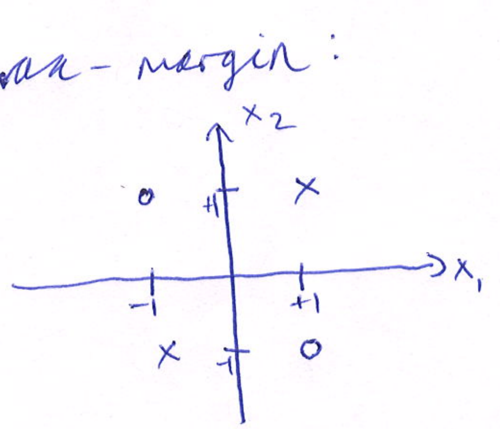
$y(\underline{x}_n) > 0.5 \Rightarrow \underline{w}^T \underline{x}_n > 0$ [if $t_n = 1$]

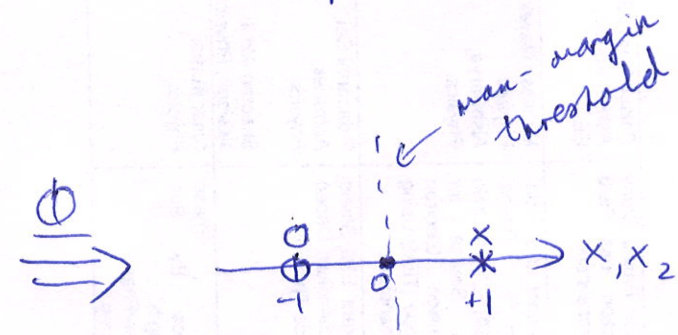
$y(\underline{x}_n) < 0.5 \Rightarrow \underline{w}^T \underline{x}_n < 0$ [if $t_n = 0$]

Clearly, the likelihood will be maximised when $y_n = t_n \;\forall n$. For y_n to be $1/0$, $\underline{w}^T \underline{x}_n$ should go to $+\infty / -\infty$. Thus, once $\underline{w}$ specifies a separating hyperplane as above, then taking $\|\underline{w}\| \to \infty$ will maximise the likelihood.

3. (a) No, not separable. (XOR problem)

Max-margin:



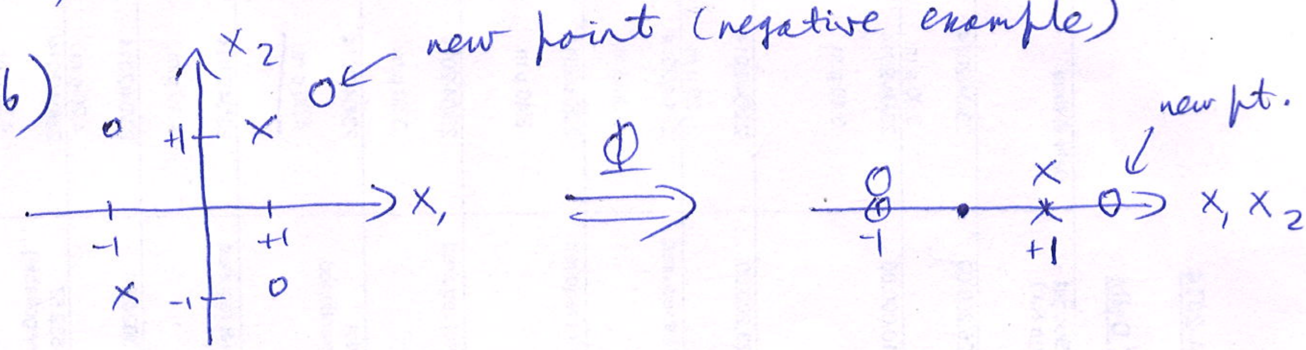


(showing just the new feature)

Only the new feature is needed, with a threshold at 0; thus, the simplest $\underline{w} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$.

Note that this $\underline{w}$ gives functional margin $= 1$, as discussed in class.

3. (b)



Clearly non-separable, even with the new feature. (Positive examples lie in-between negative examples on all three dimensions.)

(c) $K(\underline{x}, \underline{x}') = \Phi(\underline{x}) \cdot \Phi(\underline{x}')$

$$= \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ x_1 x_2 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ x_1' \\ x_2' \\ x_1' x_2' \end{pmatrix}$$

$$= 1 + x_1 x_1' + x_2 x_2' + x_1 x_1' x_2 x_2'$$

It is a dot product, and thus by definition a Mercer kernel.