Brief background on eigenvectors and eigenvalues:

A vector w is an eigenvector of matrix Σ with eigenvalue λ if and only if $\Sigma w = \lambda w$.

A size of an eigenvector is given by the magnitude of its eigenvalue.

The eigenvalues of a symmetric matrix are always real numbers and its eigenvectors form an orthonormal basis. Since the covariance matrix Σ is symmetric, its eigenvalues are real.

Principal component analysis

We have seen that the solution to the PCA objective is given by the largest eigenvector of the covariance matrix. To recap what we saw in the slides we want to solve the optimization problem below:

$$argmax_w\, w^T \Sigma w\, s.t.\ w^T w = 1$$

where Σ is the covariance matrix.

We form the Lagrangian:

$$L(w, \lambda) = w^T \Sigma w + \lambda(w^T w - 1)$$

We have to solve the Lagrangian:

$$dL/dw = 0\ and\ dL/d\lambda = 0$$

When we do $dL/dw = 0$ we get

$$dL(w, \lambda)/dw = 2\Sigma w + 2\lambda w = 0$$

$$dL(w, \lambda)/dw = 2\Sigma w =- 2\lambda w$$

which means that the solution w is just the eigenvector of Σ.

Eigenvectors of various matrices

Matrix	Meaning of the eigenvector
Covariance matrix XX^T	The data will have the highest variance when projected on the largest eigenvector
Kernel matrix X^TX	The projection of the data on the eigenvector of the covariance matrix that gives the highest variance
Data matrix X	Singular vectors are the eigenvectors of the covariance matrix and the singular values squared are the eigenvalues of the covariance matrix

---

Random projections

Random projections will preserve distances between pairs of datapoints. This may not be useful for visualization because we may need several random projections. But it will be useful for very big data with high dimensions. This can occur in text analysis where you have millions of datapoints and let’s say a 100,000 dimensions. In this case you may want to perform a random projection to let us say 100 dimensions and then analyse the big dataset in this reduced dimensionality space.

Supervised dimensionality reduction

What is a simple objective to do supervised dimensionality reduction?

In PCA we looked for a w of length 1 that maximizes the variance of the projected data. In supervised learning we have labels. Can we use the labels to do a better projection? Consider the objective below:

$$argmax_w\, w^T m_1 - w^T m_2 \;\; s.t.\; w^T w = 1$$

where $m_1 = (1/|C_1|)\Sigma_{x_{i\in C1}} x_i$ and $m_2 = (1/|C_2|)\Sigma_{x_{i\in C2}} x_i$

Let us modify the objective to be $argmax_w\ (w^T m_1 - w^T m_2)^2\ \ s.t.\ w^T w = 1$. Rewrite the objective in the form $argmax_w\ w^T M w\ \ s.t.\ w^T w = $ 1and now we can solve it just like we did PCA. But what is M?

We can find the optimal w to the above problem with Lagrange multipliers.

We can add to the above objective the variance in the denominator and now we have the Fisher or Linear Discriminant Analysis (LDA).

We can add the variance to the denominator and then search for the projection that maximizes the ratio of the mean to variance.

$$argmax_w \frac{(w^T m_1 - w^T m_2)^2}{s_1^2 + s_2^2}\ \ s.t.\ w^T w = 1$$

We can write the above ratio as

$$argmax_w \frac{w^T S_b w}{w^T S_w w}\ \ s.t.\ w^T w = 1$$

where Sb is the between class scatter matrix and Sw is the within class matrix.

If we add Sb and Sw we can get the total scatter matrix St which is also the covariance matrix that we saw earlier in PCA and in multivariate Gaussian classification.

We solve the above objective again using Lagrange multipliers like we did for PCA and that tells us that our solution w is the largest eigenvector of $S_w^{-1} S_b$.

Calculating the inverse can be a problem if the determinant of the matrix is near 0 (which can happen frequently). To avoid this we consider a different supervised dimensionality reduction called the maximum margin criterion.

$$argmax_w\ ||m_1 - m_2||^2 - s(C_1) - s(C_2)\ s.t.\ w^T w = 1$$

where s(C1) and s(C2) represents the variance of class C1 and C2 in the projected space. As we did for LDA we will use the between class scatter matrix and the within class matrix to represent the difference of means and the class variances. This gives us the objective

$$argmax_w\ S_b - S_w\ \ s.t.\ w^T w = 1$$

With Lagrange multipliers we can show that the solution we are looking for is the largest eigenvector of Sb-Sw. Note that since St=Sb+Sw we can rewrite this as

St - Sw = Sb ->
St - 2Sw = Sb - Sw
