1 Definitions

1. **Dimensionality Reduction:** The transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data. Dimensionality Reduction is useful because time and space can be saved by using fewer features, especially when dealing with a large dataset.

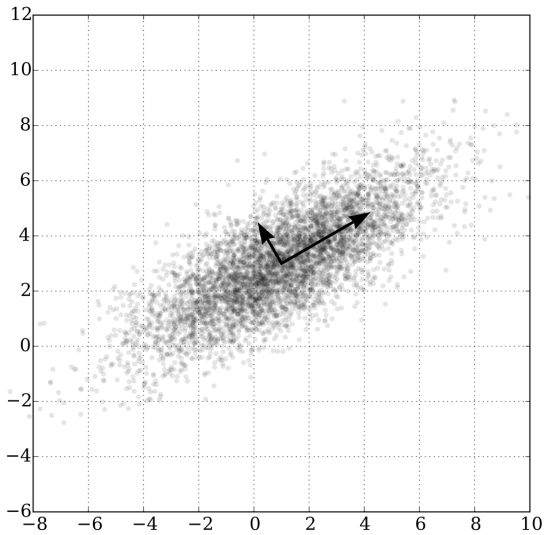
2. **Principal Component Analysis:** A dimensionality reduction method that transforms a set of features in a dataset into a smaller number of features, called principal components, while also trying to retain as much information from the original dataset as possible. The principal components are uncorrelated, and the first principal component will explain the most variance in the original dataset.

3. **Lagrange Multiplier:** A simple and elegant method of finding the local minima or local maxima of a function subject to equality or inequality constraints. It allows us to wrap the constraint into the objective function.

4. **Eigenvalue & Eigenvector:** For a matrix $A_{n\times n}$ and column vector $v_{n\times 1}$, we consider v to be an eigenvector of A and $\lambda \neq 0$ to be an eigenvalue of A iff $Av = \lambda v$. Note how for a given eigenvalue, we can have infinite eigenvectors associated with it by taking constant multiples of v. To be consistent, we are usually concerned with the eigenvector that's an unit vector (i.e. $\|v\|_2 = 1$).

5. **Eigen Value Decomposition:** If a matrix $A_{n\times n}$ has a full set of eigenvalues, then we can write $A = X\Lambda X^{-1}$ where $\Lambda_{n\times n}$ is a diagonal matrix with eigenvalues on the diagonal and $X_{n\times n}$ is the matrix whose columns are the eigenvectors corresponding to the eigenvalues in Λ.

6. **Singular Value Decomposition (SVD):** the factorization of a matrix $A_{m\times n}$ into three matrices U, W, V^T. U is defined as an $m \times n$ matrix of the orthonormal eigenvectors of AA^T. W is defined as an $n \times n$ diagonal matrix of the singular values, which are the square roots of the eigenvalues of A^TA. Finally V^T is the transpose of a $n \times n$ matrix containing the orthonormal eigenvectors of A^TA.



2 Naive Bayes with Laplace Smoothing

We're going to repeat the example from last week but introduce Laplace Smoothing as the Prior. It was very convenient to assume that the featues are independent given the class–it allowed us to simplify the calculations a lot. But, we found a problem with our example last week. Whenever a word/token didn't occur for a given class, the probability for that token given that class becomes 0, hence the probability of that class also becomes 0. This is a massive drawback. We cannot extrapolate our model to an example that contains a word that we have not seen before in our training data. If we tried, the probability for each class would just be 0.

In order to get around this, we can use something called Laplace Smoothing. We can smooth our probabilities upwards by some constant α. To be more precise,

$$P(x_i|Y=k) = \frac{\text{Number of samples that belong to class k and contains the token } x_i + \alpha}{\text{Number of samples that belong to class k } + \alpha\dot{K}}$$

K is the total number of classes and α is a constant hyperparameter. Compare this to the formula from before:

$$P(x_i|Y=k) = \frac{\text{Number of samples that belong to class k and contains the token } x_i}{\text{Number of samples that belong to class k}}$$

Let's repeat the same example from last week but use Laplace Smoothing this time to see how we can avoid getting zero as our class probability. Use $\alpha = 1$.

Consider the following training samples:

SPAM	Email Body
1	Money is free now
0	Pat teach 315
0	Pat free to teach
1	Sir money to teach
1	Pat free money now
0	Teach 315 now
0	Pat to teach 315

The vocabulary consists of the following words: {315, free, is, money, now, Pat, Sir, teach, to, tomorrow}. Compute the following probabilities:

1. Fill in the tables below:

$P(Y=1)$	$P(Y=0)$

$P(Y=1) = 3/7 \quad P(Y=0) = 4/7$

j	$P(X_j = 1 \mid Y = 1)$	$P(X_j = 1 \mid Y = 0)$
315	1/5	4/6
free	3/5	2/6
is	2/5	1/6
money	4/5	1/6
now	3/5	2/6
Pat	2/5	4/6
Sir	2/5	1/6
teach	2/5	5/6
to	2/5	3/6
tomorrow	1/5	1/6

2. Consider the following email body: X = Pat teach now. Accounting for the laplace smoothing, fill in the table below:

j	x	$P(X_j = x \mid Y = 1)$	$P(X_j = x \mid Y = 0)$
315	0	4/5	2/6
free	0	2/5	4/6
is	0	3/5	5/6
money	0	1/5	5/6
now	1	3/5	2/6
Pat	1	2/5	4/6
Sir	0	3/5	5/6
teach	1	2/5	5/6
to	0	3/5	3/6
tomorrow	0	4/5	5/6

3. Reminder that with naive Bayes, $P(Y, X_1, \ldots, X_M) = \prod P(X \mid Y)P(Y)$.

$P(Y = 1, X_1, ..., X_M)$	$P(Y = 0, X_1, ..., X_M)$

$P(Y = 1, X_1, ..., X_m) = \prod P(X \mid Y = 1)P(Y = 1)$
4/5 * 2/5 * 3/5 * 1/5 * 3/5 * 2/5 * 3/5 * 2/5 * 3/5 * 4/5 * 3/7
= 0.00045500708

$P(Y = 0, X_1, ..., X_M)$
2/6 * 4/6 * 5/6 * 5/6 * 2/6 * 4/6 * 5/6 * 5/6 * 3/6 * 3/6 * 4/7
=0.00340213817

4.

$P(Y = 1 \mid X_1, ..., X_M)$	$P(Y = 0 \mid X_1, ..., X_M)$
0.00045500708 / (0.00045500708 + 0.00340213817) = 0.118	0.00340213817 / (0.00045500708 + 0.00340213817) = 0.882

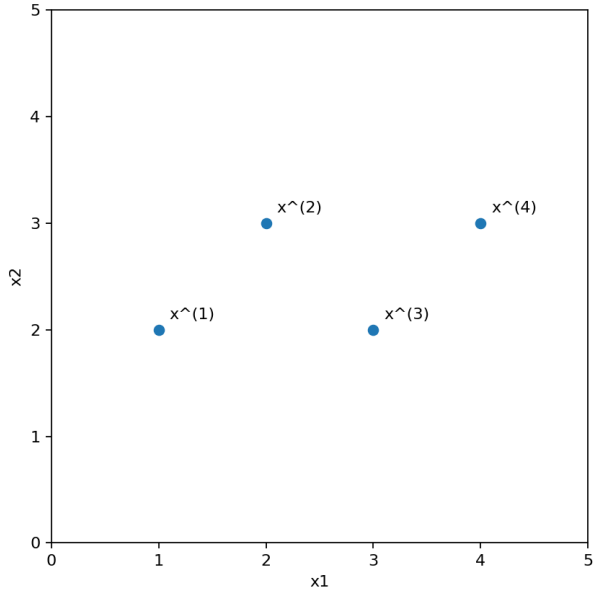
We see that $P(Y = 1 \mid X_1, ..., X_M)$ is no longer 0.

What's the effect of α? Compare when $\alpha = 1$ with when $\alpha = \infty$

As α increases, the likelihood probability moves towards uniform distribution.

3 PCA Walkthrough

Consider dataset $\mathcal{D} = \{\mathbf{x}^{(1)} = \begin{bmatrix}1\\2\end{bmatrix}, \mathbf{x}^{(2)} = \begin{bmatrix}2\\3\end{bmatrix}, \mathbf{x}^{(3)} = \begin{bmatrix}3\\2\end{bmatrix}, \mathbf{x}^{(4)} = \begin{bmatrix}4\\3\end{bmatrix}\}$. A visualization of the dataset is as below.



1. Centering is crucial for PCA. We must preprocess data so that all features have zero mean before applying PCA, i.e.

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^{(i)} = \vec{0}$$

Compute the centered dataset $\mathcal{D}' = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}\}$.

First, note that $\mathbb{E}[x_1] = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_1^{(i)} = 2.5$ and $\mathbb{E}[x_2] = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_2^{(i)} = 2.5$.

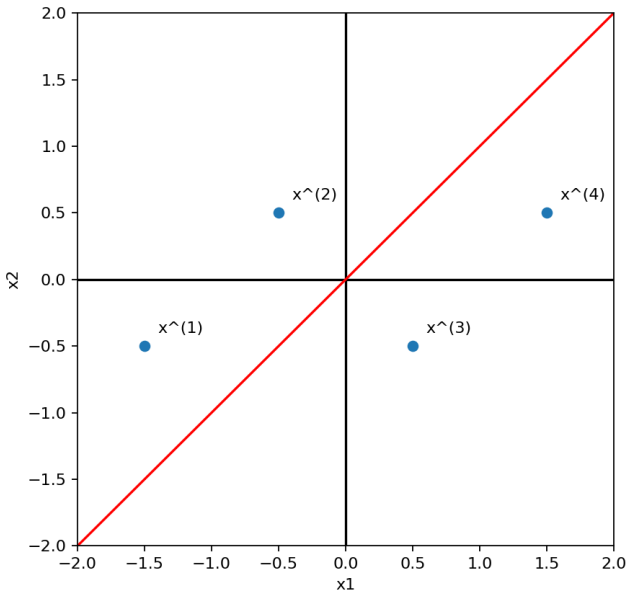
$$\mathbf{x}^{(1)} = \begin{bmatrix}-1.5\\-0.5\end{bmatrix} \qquad \mathbf{x}^{(2)} = \begin{bmatrix}-0.5\\0.5\end{bmatrix}$$
$$\mathbf{x}^{(3)} = \begin{bmatrix}0.5\\-0.5\end{bmatrix} \qquad \mathbf{x}^{(4)} = \begin{bmatrix}1.5\\0.5\end{bmatrix}$$

2. In order to easily compute the projected coordinates of data, we need to make the projected directions unit vectors. Suppose we want to project our data onto the vector $\mathbf{v} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Normalize $\mathbf{v}$ to be a unit vector.

   To make $\mathbf{v}$ into a unit vector, we divide $\mathbf{v}$ by its magnitude. The magnitude of a vector is given by the L2 norm.

   $$||\mathbf{v}|| = \sqrt{1^2 + 1^2} = \sqrt{2}$$
   $$\mathbf{v} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

3. The centered data should now look like the following:



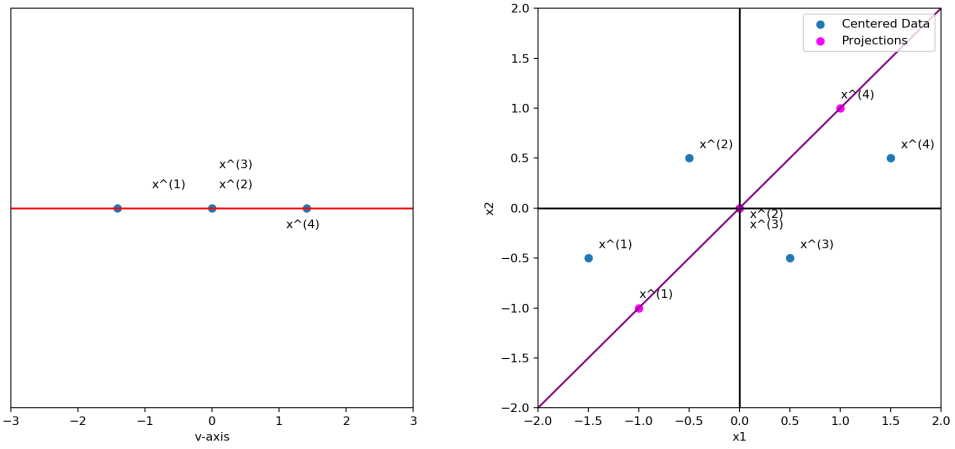
   Suppose we want to project the centered data onto $\mathbf{v}$, where $\mathbf{v}$ goes through the origin. Compute the magnitude of the projections, i.e. compute $z^{(i)} = \mathbf{v}^T\mathbf{x}^{(i)}, \forall 1 \leq i \leq N$.

$$z^{(1)} = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}] \begin{bmatrix} -1.5 \\ -0.5 \end{bmatrix} = \frac{-3}{2\sqrt{2}} - \frac{1}{2\sqrt{2}} = -\frac{4}{2\sqrt{2}} = -\sqrt{2}$$
$$z^{(2)} = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}] \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix} = \frac{-1}{2\sqrt{2}} + \frac{1}{2\sqrt{2}} = 0$$
$$z^{(3)} = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}] \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix} = \frac{-1}{2\sqrt{2}} + \frac{1}{2\sqrt{2}} = 0$$
$$z^{(4)} = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}] \begin{bmatrix} 1.5 \\ 0.5 \end{bmatrix} = \frac{3}{2\sqrt{2}} + \frac{1}{2\sqrt{2}} = \frac{4}{2\sqrt{2}} = \sqrt{2}$$

4. Let $\mathbf{x}^{(i)'}$ be the projected point of $\mathbf{x}^{(i)}$. Note that $\mathbf{x}^{(i)'} = \mathbf{v}^T\mathbf{x}^{(i)}\mathbf{v} = z^{(i)}\mathbf{v}$. Compute the projected coordinates $\mathbf{x}^{(1)'}, \mathbf{x}^{(2)'}, \mathbf{x}^{(3)'}$, and $\mathbf{x}^{(4)'}$.

$$\mathbf{x}^{(1)'} = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \qquad \mathbf{x}^{(2)'} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$
$$\mathbf{x}^{(3)'} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \mathbf{x}^{(4)'} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Below is a visualization of the projections:



5. One of the two goals of PCA is to find new directions to project our dataset onto such that it **minimizes the reconstruction error**, where the reconstruction error is defined as following:

$$\text{Reconstruction Error} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}^{(i)'} - \mathbf{x}^{(i)}\|_2^2$$

What is the reconstruction error in our case?

$$\begin{aligned}
\text{Recon. Error} &= \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}^{(i)'} - \mathbf{x}^{(i)}\|_2^2 \\
&= \frac{1}{N}(\| \begin{bmatrix}-1\\-1\end{bmatrix} - \begin{bmatrix}-1.5\\-0.5\end{bmatrix} \|_2^2 + \| \begin{bmatrix}0\\0\end{bmatrix} - \begin{bmatrix}-0.5\\0.5\end{bmatrix} \|_2^2 + \| \begin{bmatrix}0\\0\end{bmatrix} - \begin{bmatrix}0.5\\-0.5\end{bmatrix} \|_2^2 + \| \begin{bmatrix}1\\1\end{bmatrix} - \begin{bmatrix}1.5\\0.5\end{bmatrix} \|_2^2) \\
&= \frac{1}{4}((\frac{1}{2})^2 + (-\frac{1}{2})^2 + (\frac{1}{2})^2 + (-\frac{1}{2})^2 + (-\frac{1}{2})^2 + (\frac{1}{2})^2 + (-\frac{1}{2})^2 + (\frac{1}{2})^2) \\
&= \frac{1}{4}(\frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}) \\
&= \frac{1}{2}
\end{aligned}$$

6. Another goal is to find new directions to project our dataset onto such that it **maximizes the variance of the projections**, where the variance of projections is defined as following:

$$\begin{aligned}
\text{variance of projection} &= \frac{1}{N}\sum_{i=1}^{N}(z^{(i)} - \mathbb{E}[z])^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}(\mathbf{v}^T\mathbf{x}^{(i)} - \frac{1}{N}\sum_{j=1}^{N}\mathbf{v}^T\mathbf{x}^{(j)})^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}(\mathbf{v}^T\mathbf{x}^{(i)} - \mathbf{v}^T(\frac{1}{N}\sum_{j=1}^{N}\mathbf{x}^{(j)}))^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}(\mathbf{v}^T\mathbf{x}^{(i)} - \mathbf{v}^T\vec{0})^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}(\mathbf{v}^T\mathbf{x}^{(i)})^2
\end{aligned}$$

What is the variance of the projections?

$$\begin{aligned}\text{variance} &= \frac{1}{N}\sum_{i=1}^{N}(z^{(i)})^2 \\ &= \frac{1}{4}((-\sqrt{2})^2 + 0^2 + 0^2 + (\sqrt{2})^2) \\ &= \frac{1}{4}(2+0+0+2) \\ &= 1\end{aligned}$$

7. What methods can we use to find the principal components?

Eigenvalue decomposition and SVD

8. What is the first principal component of $X = [x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}]^T$ using the eigenvalue decomposition?

$$X = \begin{bmatrix} -1.5 & -0.5 \\ -0.5 & 0.5 \\ 0.5 & -0.5 \\ 1.5 & 0.5 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 5 & 1 \\ 1 & 1 \end{bmatrix}$$

Solve the eigenvalues and eigenvectors for $X^T X$.

$$X^T X v_i = \lambda_i v_i$$

$$\lambda_1 = 3 + \sqrt{5}, \quad v_1 = \begin{bmatrix} 2+\sqrt{5} \\ 1 \end{bmatrix}$$

$$\lambda_2 = 3 - \sqrt{5}, \quad v_2 = \begin{bmatrix} 2-\sqrt{5} \\ 1 \end{bmatrix}$$

So the first principal component is $v_1/||v_1||_2 = \begin{bmatrix} 0.973 \\ 0.230 \end{bmatrix}$.

4 Fun With Lagrange and PCA

Consider the following optimization where A is a symmetric matrix:

$$\begin{aligned} \hat{\mathbf{v}} = \underset{\mathbf{v}}{\text{argmax.}} \quad & \mathbf{v}^T A \mathbf{v} \\ \text{s.t.} \quad & \|\mathbf{v}\|_2^2 = 1 \end{aligned}$$

1. Formulate the Lagrangian, $\mathcal{L}(\mathbf{v}, \lambda)$.

$$\mathcal{L}(\mathbf{v}, \lambda) = \mathbf{v}^T A \mathbf{v} - \lambda(\|\mathbf{v}\|_2^2 - 1)$$

2. Write the gradient of the Lagrangian with respect to $\mathbf{v}$, $\nabla_{\mathbf{v}} \mathcal{L}(\mathbf{v}, \lambda)$.

$$\mathcal{L}(\boldsymbol{\theta}, \lambda) = \mathbf{v}^T A \mathbf{v} - \lambda(\mathbf{v}^T \mathbf{v} - 1)$$

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \lambda) = 2A\mathbf{v} - 2\lambda\mathbf{v}$$

3. After setting $\nabla_{\mathbf{v}} \mathcal{L}(\mathbf{v}, \lambda)$ equal to zero, what can you say about the relationship between λ, $\mathbf{v}$, and the eigenvalues and eigenvectors of A?

$$\begin{aligned} 2A\mathbf{v} - 2\lambda\mathbf{v} &= 0 \\ A\mathbf{v} &= \lambda\mathbf{v} \end{aligned}$$

$\mathbf{v}$ is an eigenvector of A and λ is the corresponding eigenvalue of A!