# CS/ECE/ME 532
# Homework 3: Subspaces and orthogonality

1. **Orthogonal columns**. Consider the matrix and vector

$$\boldsymbol{A} = \begin{bmatrix} 3 & 1 \\ 0 & 3 \\ 0 & 4 \end{bmatrix} \quad \text{and} \quad \boldsymbol{b} = \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix}.$$

**a)** By hand, find two orthonormal vectors that span the plane spanned by columns of $\boldsymbol{A}$.

**SOLUTION:** We can begin by normalizing the first vector. This gives us:

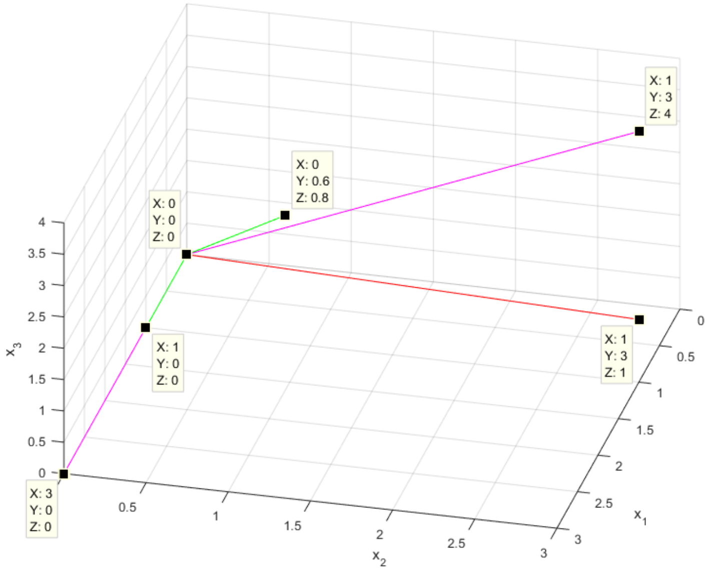
$$\boldsymbol{u}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

To find the second vector, we need a vector that is orthogonal to $\boldsymbol{u}_1$ and in the span of the columns of $\boldsymbol{A}$. One way to ensure orthogonality in this case is to cancel out the first component of the second column of $\boldsymbol{A}$. So subtract $\boldsymbol{u}_1$ from $\boldsymbol{a}_{\cdot 2}$, and normalize:

$$\boldsymbol{u}_2 = \begin{bmatrix} 0 \\ \frac{3}{5} \\ \frac{4}{5} \end{bmatrix}$$

**b)** Make a sketch of these vectors and the columns of $\boldsymbol{A}$ in three dimensions.

**SOLUTION:** Here is a sketch. Again, we will accept anything that is roughly correct. $\boldsymbol{A}$-vectors in magenta, $\boldsymbol{u}$-vectors in green, $\boldsymbol{b}$-vector in red.



**c)** Use these vectors to compute the LS estimate $\widehat{\boldsymbol{b}} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}$.

**SOLUTION:** knowing the orthogonalization of $\boldsymbol{A}$ makes the computation much easier! The least-squares solution $\widehat{\boldsymbol{b}}$ will be the same if we replace $\boldsymbol{A}$ by $\boldsymbol{U}$. Then, we have:

$$\begin{aligned}
\widehat{\boldsymbol{b}} &= \boldsymbol{U}(\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{b} \\
&= \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{b} \\
&= \begin{bmatrix} 1 & 0 \\ 0 & \frac{3}{5} \\ 0 & \frac{4}{5} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{3}{5} \\ 0 & \frac{4}{5} \end{bmatrix}^T \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 \\ 0 & \frac{3}{5} \\ 0 & \frac{4}{5} \end{bmatrix} \begin{bmatrix} 1 \\ \frac{13}{5} \end{bmatrix} \\
&= \begin{bmatrix} 1 \\ \frac{39}{25} \\ \frac{52}{25} \end{bmatrix}
\end{aligned}$$

Where we used the fact that $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ because $\boldsymbol{U}$ is orthogonal. We can also check that the answer is the same using $\boldsymbol{A}$ (but more work):

$$\begin{aligned}
\widehat{\boldsymbol{b}} &= \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b} \\
&= \begin{bmatrix} 3 & 1 \\ 0 & 3 \\ 0 & 4 \end{bmatrix} \left( \begin{bmatrix} 3 & 1 \\ 0 & 3 \\ 0 & 4 \end{bmatrix}^T \begin{bmatrix} 3 & 1 \\ 0 & 3 \\ 0 & 4 \end{bmatrix} \right)^{-1} \begin{bmatrix} 3 & 1 \\ 0 & 3 \\ 0 & 4 \end{bmatrix}^T \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} 3 & 1 \\ 0 & 3 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} 9 & 3 \\ 3 & 26 \end{bmatrix}^{-1} \begin{bmatrix} 3 \\ 14 \end{bmatrix} \\
&= \frac{1}{225} \begin{bmatrix} 3 & 1 \\ 0 & 3 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} 26 & -3 \\ -3 & 9 \end{bmatrix} \begin{bmatrix} 3 \\ 14 \end{bmatrix} \\
&= \frac{1}{225} \begin{bmatrix} 3 & 1 \\ 0 & 3 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} 36 \\ 117 \end{bmatrix} \\
&= \frac{1}{225} \begin{bmatrix} 225 \\ 351 \\ 468 \end{bmatrix} \\
&= \begin{bmatrix} 1 \\ \frac{39}{25} \\ \frac{52}{25} \end{bmatrix}
\end{aligned}$$

**2. Tikhonov regularization**. Sometimes we have competing objectives. For example, we want to find an $x$ that minimizes $\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2$ (least-squares), but we also want the weights $\boldsymbol{x}$ to be small. One way to achieve a compromise is to solve the following problem:

$$\text{minimize} \quad \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\,\|\boldsymbol{x}\|_2^2 \tag{1}$$

where $\lambda > 0$ is a parameter we choose that determines the relative weight we want to assign to each objective. This is called *Tikhonov regularization* (also known as $L_2$ regularization).

**a)** Solve the optimization problem (1) by finding an expression for the minimizer $\widehat{\boldsymbol{x}}$.
**Hint**: one approach is to reformulate (1) as a modified least-squares problem with different "$\boldsymbol{A}$" and "$\boldsymbol{b}$" matrices. Another approach is to use the vector derivative method we saw in class.

**SOLUTION:** the augmented cost function is an ordinary least-squares problem in disguise. To see why, notice that

$$\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_2^2 = \left\| \begin{bmatrix} \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \\ \sqrt{\lambda}\boldsymbol{x} \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{A} \\ \sqrt{\lambda}\boldsymbol{I} \end{bmatrix} \right\|_2^2$$

Applying the least-squares formula to these new matrices, we find:

$$\begin{aligned} \widehat{\boldsymbol{x}} &= \left( \begin{bmatrix} \boldsymbol{A} \\ \sqrt{\lambda}\boldsymbol{I} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{A} \\ \sqrt{\lambda}\boldsymbol{I} \end{bmatrix} \right)^{-1} \begin{bmatrix} \boldsymbol{A} \\ \sqrt{\lambda}\boldsymbol{I} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{0} \end{bmatrix} \\ &= \left(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I}\right)^{-1} \boldsymbol{A}^T\boldsymbol{b} \end{aligned}$$

Alternatively, we can find the modified normal equations directly by differentiating the cost function. Doing so, we obtain:

$$\begin{aligned} \frac{d}{d\boldsymbol{x}}\left(\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_2^2\right) &= \frac{d}{d\boldsymbol{x}}\left((\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})^T(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) + \lambda\boldsymbol{x}^T\boldsymbol{x}\right) \\ &= \frac{d}{d\boldsymbol{x}}\left(\boldsymbol{x}^T(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})\boldsymbol{x} - 2\boldsymbol{b}^T\boldsymbol{A}\boldsymbol{x}\right) \\ &= 2(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I}) - 2\boldsymbol{A}^T\boldsymbol{b} \end{aligned}$$

Setting the derivative equal to zero, we obtain $\widehat{\boldsymbol{x}} = \left(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I}\right)^{-1} \boldsymbol{A}^T\boldsymbol{b}$ as before.

**b)** Suppose that $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, with $m < n$. Is there a unique least squares solution? Is there a unique solution to (1) ? Explain your answers.

**SOLUTION:** There is always a unique solution to this problem, regardless of the dimensions of $\boldsymbol{A}$. There are many ways to prove this. One way is to examine the modified least-square problem from part (a). The solution will be unique as long as the modified $\boldsymbol{A}$ matrix has full column rank. In other words:

$$\text{does} \quad \widehat{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{A} \\ \sqrt{\lambda}\boldsymbol{I} \end{bmatrix} \quad \text{have full column rank?}$$

Again, many ways to see why the answer is yes. If $\widehat{\boldsymbol{A}}\boldsymbol{x} = 0$, then we have $\boldsymbol{A}\boldsymbol{x} = 0$ (first block) and $\sqrt{\lambda}\boldsymbol{x} = 0$ (second block). Therefore $\boldsymbol{x} = 0$, so $\widehat{\boldsymbol{A}}$ has linearly independent columns.

Another way to prove this is to recall from class that $\boldsymbol{A}^T\boldsymbol{A} \succeq 0$ (positive semidefinite). Also, we have $\lambda\boldsymbol{I} \succ 0$ (positive definite). Therefore, $(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I}) \succ 0$. Positive-definite matrices are always invertible, and so we are done.

**3. Gram-Schmidt**. Write your own code to perform Gram-Schmidt orthogonalization. Your code should take as input a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and return as output a matrix $\boldsymbol{U} \in \mathbb{R}^{m\times r}$ where $\boldsymbol{U}$ is orthogonal and has the same range as $\boldsymbol{A}$. Note that $r$ will indicate the rank of $\boldsymbol{A}$, so your code can also be used to find the rank of a matrix!

**SOLUTION:** Here is a (spectacularly short) solution in Matlab:

```
function U = gram_schmidt( A )
   % GRAM-SCHMIDT CODE
   %  U = gram_schmidt( A )
   %  where U'*U=I and range(U)=range(A)
   %  number of columns of U is the rank of A.

   [m,n] = size(A);
   U = zeros(m,0);            % start with empty matrix
   for i = 1:n
      v = A(:,i);             % the current column of A
      v = v - U*(U'*v);       % project onto current output set
      if norm(v) > 1e-12      % ensure linear independence
         U = [U v/norm(v)]; % normalize and add to the set
      end
   end
end  % end of function
```

**a)** Test your code by applying it to Problem 1 above.

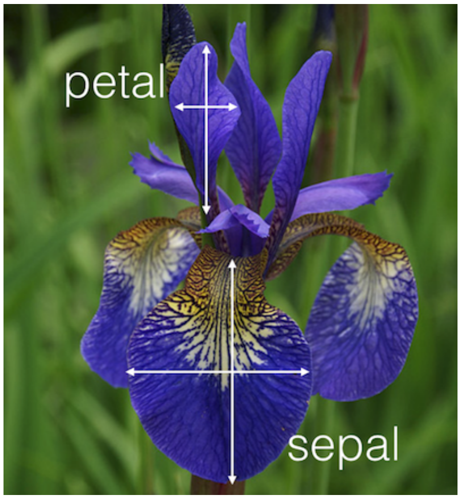
**SOLUTION:** Result is the same as the solution computed to problem 1, as expected.

**b)** Use your code to determine the rank of the following matrices and compare the result to Matlab's `rank` function (or Python's `numpy.linalg.matrix_rank` function).

$$\boldsymbol{A}_1 = \begin{bmatrix} 3 & 1 & 2 \\ 0 & 3 & 3 \\ 0 & 4 & 4 \\ 6 & 1 & 4 \end{bmatrix} \qquad \boldsymbol{A}_2 = \begin{bmatrix} 1 & 1 & 2 \\ 0 & 3 & 3 \\ 0 & 4 & 4 \\ 3 & 1 & 4 \end{bmatrix}$$

**SOLUTION:** Applying the code to $\boldsymbol{A}_1$ results in a $4 \times 3$ matrix (rank 3). Applying the code to $\boldsymbol{A}_2$ results in a $4 \times 2$ matrix (rank 2).

**4.** In 1936 Ronald Fisher published a famous paper on classification titled "The use of multiple measurements in taxonomic problems." In the paper, Fisher study the problem of classifying iris flowers based on measurements of the sepal and petal widths and lengths, depicted in the image below.



Fisher's dataset is available in Matlab (`fisheriris.mat`) and is widely available on the web (e.g., Wikipedia). The dataset consists of 50 examples of three types of iris flowers. The sepal and petal measurements can be used to classify the examples into the three types of flowers.

**a)** Formulate the classification task as a least squares problem. Least squares will produce real-valued predictions, not discrete labels or categories. What might you do to address this issue?

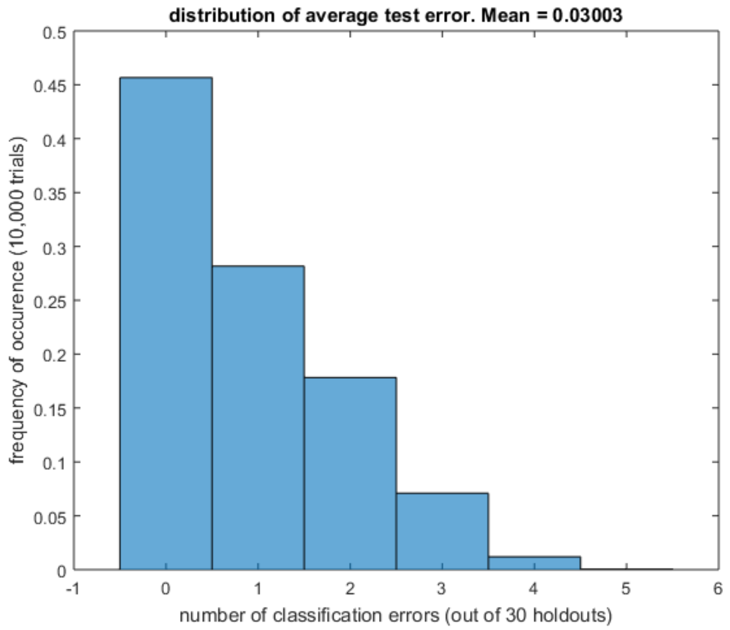
**SOLUTION:** One possibility is to assign numerical values to the labels. For example: setsoa $= -1$, versicolor $= 0$, virginica $= 1$. Then it becomes standard least-squares.

**b)** Write a Matlab or Python program to "train" a classifier using LS based on 40 labeled examples of each of the three flower types, and then test the performance of your classifier using the remaining 10 examples from each type. Repeat this with many different randomly chosen subsets of training and test. What is the average test error (number of mistakes divided by 30)?

**SOLUTION:** The code below finds a random training set of size 40 (for each label), computes the classifier (using the method of part a), computes the error on the holdout set of 10 (for each label), then finds the test error. This is repeated 10,000 times and the average test error is recorded. I also plotted the error distribution. The holdout set has 30 elements, and random classifiers have 0/30 error roughly 45% of the time and 1/30 error roughly 27% of the time. Classifiers that make 4/30 or more errors are very rare. Here is the code:

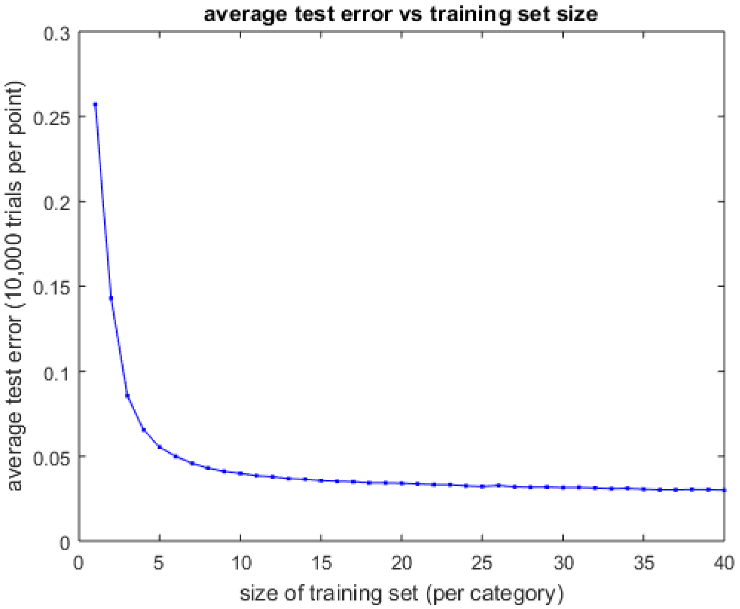
```
load fisheriris
A = meas;
b = kron( [-1; 0; 1], ones(50,1) );

N = 10000;               % number of random trials
errs = zeros(N,1);      % where we store error values
num_train = 40;         % size of training set

for i = 1:N

   % randomly pick training and holdout sets
   r = randperm(50);
   r = r(1:num_train);              % random set for training
   rc = setdiff(1:50,r);            % remaining are the holdouts
   train = [r r+50 r+100];          % training set
   holdout = [rc rc+50 rc+100];     % holdout set

   % train the classifier
   At = A(train,:);
   bt = b(train,:);
   xt = inv(At'*At)*At'*bt;

   % use classifier on holdout set
   Ah = A(holdout,:);
   bh = b(holdout,:);
   bhat = Ah*xt;

   % apply rounding to find labels
   nh = numel(holdout);
   for j = 1:nh
      if abs(bhat(j)) < 0.5
         bhat(j) = 0;
      else
         bhat(j) = sign(bhat(j));
      end
   end
   errs(i) = mean(bhat~=bh);
end

% plot histogram and mean value
avg_error = mean(errs)
figure(1); clf
histogram(errs*nh,'BinMethod','integers','Normalization','probability')
xlabel('number of classification error (out of 30 holdouts)')
ylabel('frequency of occurence (10,000 trials)')
title(['distribution of average test error. Mean = ' num2str(avg_error)])
```

and here is a plot of the histogram produced by the code. The average test error over 10,000 trials was 3%.



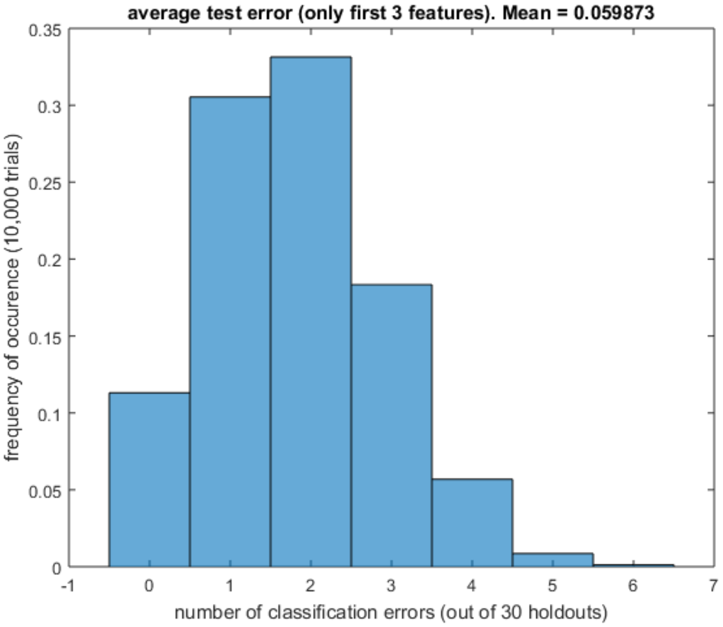
**c)** Experiment with even smaller sized training sets. Clearly we need at least one training example from each type of flower. Make a plot of average test error as a function of training set size.

**SOLUTION:** The code for this part is very similar to part (b) so we omit it. When we reduce the size of the training set, we find that the average error remains relatively constant down to about 10 samples per category (flower). This amounts to roughly 20% of the total data. Any less, and the error goes up dramatically. As we might expect, more training data means a better classifier!



**d)** Now design a classifier using only the first three measurements (sepal length, sepal width, and petal length). What is the average test error in this case?

**SOLUTION:** By slightly modifying the code of part (b), (changing `meas` to `meas(:,1:3)`) we can compute the average error that results from using only the first three features. The new histogram is shown below. This time, the error is about 5.9%.



**e)** Use a 3d scatter plot to visualize the measurements in (d). Can you find a 2-dimensional subspace that the data approximately lie in? You can do this by rotating the plot and looking for plane that approximately contains the data points.

**SOLUTION:** Upon making the scatter plot (shown below), it is clear that the points corresponding to different species are clustered (we will see how to address clustering directly later in the class!). It is also clear that there is a lower-dimensional subspace that roughly contains the data. In this case, it is a 2D hyperplane.

To estimate the orientation of this plane (and compute a projection), we need to find a basis that spans the subspace. One approximate way to do this is to pick three representative points (see labels on the figure). The points I chose were:

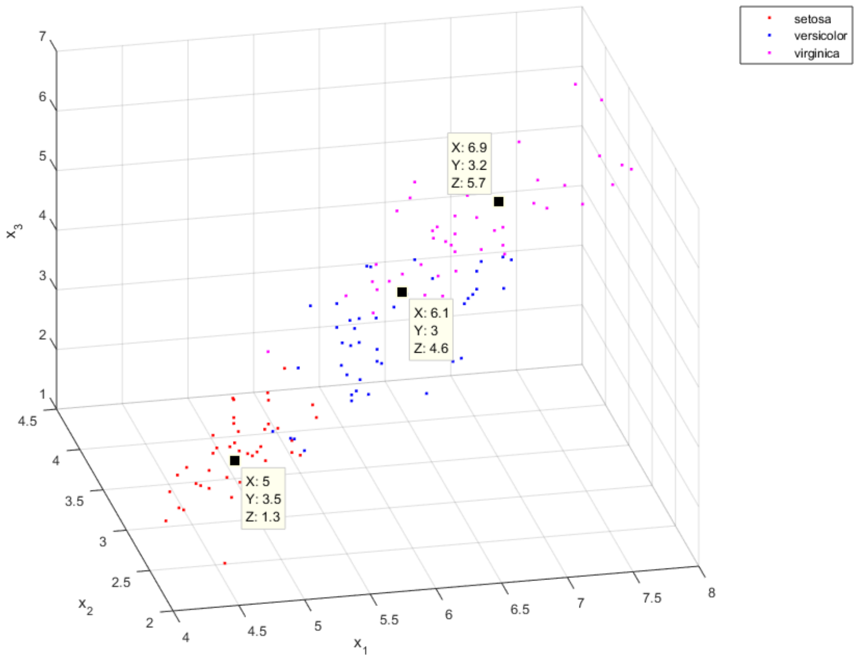
$$\boldsymbol{v}_1 = \begin{bmatrix} 5 \\ 3.5 \\ 1.3 \end{bmatrix} \qquad \boldsymbol{v}_2 = \begin{bmatrix} 6.1 \\ 3 \\ 4.6 \end{bmatrix} \qquad \boldsymbol{v}_3 = \begin{bmatrix} 6.9 \\ 3.2 \\ 5.7 \end{bmatrix}$$

Using the first vector as a reference, we can compute a basis:

$$\boldsymbol{u}_1 = \boldsymbol{v}_2 - \boldsymbol{v}_1 = \begin{bmatrix} 1.1 \\ -0.5 \\ 3.3 \end{bmatrix} \qquad \boldsymbol{u}_2 = \boldsymbol{v}_3 - \boldsymbol{v}_1 = \begin{bmatrix} 1.9 \\ -0.2 \\ 4.4 \end{bmatrix}$$

We can therefore conclude that the points are roughly aligned with the subspace $\mathcal{S}$, where

$$\mathcal{S} = \operatorname{span}\left( \begin{bmatrix} 1.1 \\ -0.5 \\ 3.3 \end{bmatrix}, \begin{bmatrix} 1.9 \\ -0.2 \\ 4.4 \end{bmatrix} \right)$$



The intuition here is that if we projected our data onto this subspace, we wouldn't lose much information. Although it is of little consequence for this example, you might imagine a scenario where the data is in $\mathbb{R}^{10^6}$ and there is a great computational benefit to finding a low-dimensional representation of the data.

**f)** Use this subspace to find a 2-dimensional classification rule. What is the average test error in this case?

**SOLUTION:** We will use least-squares again, but this time with the projected data. To project our data onto the two-dimensional subspace, let's begin by finding an orthonormal basis. We can use our Gram-Schmidt code for this, and apply it to the subspace found in part (e). The result is:

$$\boldsymbol{U} = \begin{bmatrix} 0.3130 & 0.6503 \\ -0.1423 & 0.7527 \\ 0.9390 & -0.1027 \end{bmatrix}$$

Therefore, if we are given a feature vector $\boldsymbol{v} \in \mathbb{R}^3$, its coordinates in the projected space are $\tilde{\boldsymbol{v}} = \boldsymbol{U}^T \boldsymbol{v} \in \mathbb{R}^2$. Since the feature vectors are stored as rows of $\boldsymbol{A}$, the relevant transformation is $\tilde{\boldsymbol{v}}^T = \boldsymbol{v}^T\boldsymbol{U}$. So our new $\boldsymbol{A}$ matrix (with projected features) is therefore $\boldsymbol{AU} \in \mathbb{R}^{150\times 2}$. To train a classifier on the projected features, we must solve the least-squares problem:

$$\text{minimize} \quad \|\boldsymbol{AUw} - \boldsymbol{b}\|_2^2$$

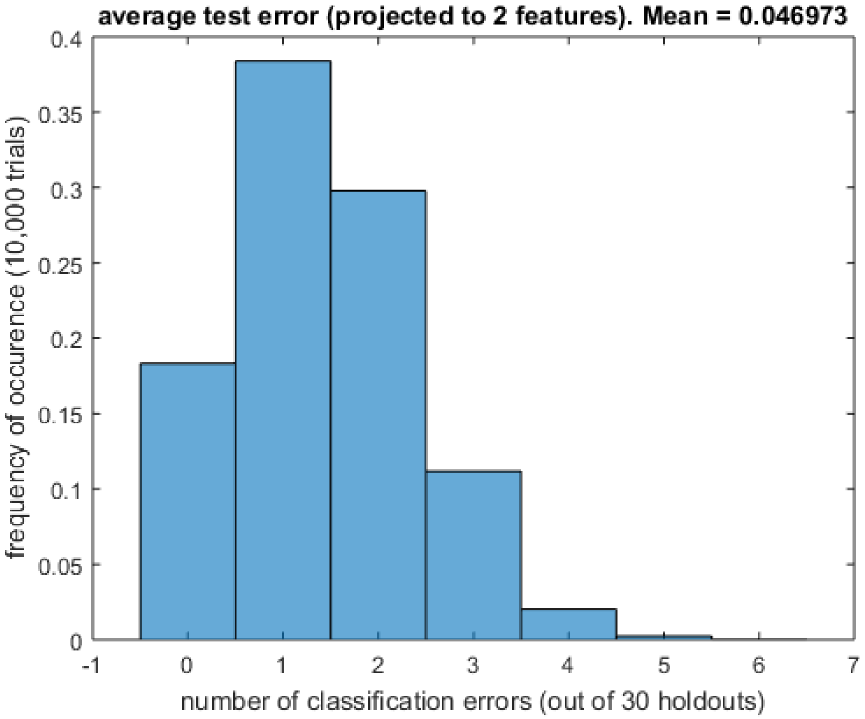
Once we have solved this to find the weight vector $\widehat{\boldsymbol{w}}$, the classification procedure is:

i. Take the vector of 3 features $\boldsymbol{v}^T$ and compute its projection $\tilde{\boldsymbol{v}}^T = \boldsymbol{v}^T\boldsymbol{U}$.

ii. Multiply the projected feature vector by the optimal weights: $s = \tilde{\boldsymbol{v}}^T\widehat{\boldsymbol{w}}$

iii. Threshold. Depending on whether $s$ is closer to $-1$, 0, or 1, assign it the label *setosa*, *versicolor*, or *virginica*.

This procedure is actually equivalent to just replacing $\boldsymbol{A}$ with $\boldsymbol{AU}$ and proceeding as in part (d). I re-ran the code from part (d) with this modification and I obtained the following error distribution:



In contrast with part (d), the error has *dropped*! We are now at 4.7% instead of the 5.9% we obtained in part (e) when we used three features.

It is worth pointing out that this sort of reduction in error is not typical. It's actually easy to prove that for any orthogonal matrix $\boldsymbol{U}$, we have $\min_{\boldsymbol{x}} \|\boldsymbol{Ax}-\boldsymbol{b}\|_2 \le \min_{\boldsymbol{w}} \|\boldsymbol{AUw}-\boldsymbol{b}\|_2$. Therefore, projecting your data onto a lower dimensional subspace before solving the least-squares problem can ***never*** improve the residual. So what is going on here? Our classifier is not actually measuring the residual! We are measuring the *error rate*, so there is an additional thresholding operation that occurs (when we round the values to $-1$, 0, or 1, depending on which is closer). As it turns out, it's possible that projecting the data can cause the error rate to decrease, even though the residual will increase.