1. (Clustering, 10 points) Clustering the Iris dataset. The dataset is included in scikit-learn (link). Treat the data $X$ with four features as unlabeled data, and cluster it into three clusters with k-Means (link). Then calculate the normalized mutual information (link) between your cluster labels and the true class labels $Y$. We know that k-Means result is affected by initialization of cluster centers, so you need to run 5 times for each setting below and choose the best performance. To make your result repeatable, you need to set the argument random_state. Set 5 different values to random_state to get 5 random runs. You can always repeat your 5 runs with these values.

   (i) (5 points) Compare the two initialization methods, init='random' and init='k-means++'. Show NMI scores you get from the two methods.

   (ii) (5 points) Fix init='k-means++', and then scale up the second feature by 10 ($X[:, 1] = X[:, 1] \times 10$), then cluster again and calculate the NMI. Do you get better or worse NMI? Explain why.

2. (Clustering, 10 points, OPTIONAL) Suppose we run k-Means and get $k$ clusters with centers $\mu_1, \ldots, \mu_k$. Now we turn this clustering problem into a $k$-class classification problem by assign a label to each cluster. Can you derive a one-vs-one multiclass classifier with linear base classifiers for the problem such that your classifier achieves accuracy 1.0?

3. (Collaborative Filtering, 10 points) In one recommendation application, users can rate items with integers from 1 to 5. Suppose we have collected 5 ratings from 3 users for 3 items.

<table>
<thead>
<tr>
<th></th>
<th>item 1</th>
<th>item 2</th>
<th>item 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>user 1</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>user 2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>user 3</td>
<td></td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

   (i) (5 points) Can you fill up missing ratings with SVD decomposition, $M = UV^T$, with $U, V \in \mathbb{R}^{3 \times 1}$? Please show your answers of $U, V$, and ratings to fill up. You don’t need to subtract the average from either rows or columns of the matrix. You don’t need to consider regularization either.

   (ii) (5 points) This method is certainly not ideal for this application. Discuss two possible issues you think most significant.

4. (Generative Model, 10 points) Suppose we have three documents

   - Machine learning is interesting
   - Ice-cream is delicious
   - I eat ice-cream when I study machine learning

Now let’s model the three documents with LDA with $K = 2$ topics. In the original LDA model, we need to infer document topics $\theta$ as well as word distributions $\beta$. In the following problem, you are given the $\beta$ matrix as below.
(i) (5 points) The topic vector $\theta$ for each document is a probability vector with two elements. Can you make your best guess of $\theta$ vectors for the three document above? Justify your guess in the sense of MLE by answering the question: why your guesses give large log-likelihood to the data? You don’t need rigorous calculation here.

(ii) (5 points) Can you randomly generate a document of length 3 with a topic vector $\theta = [0.7, 0.3]$. Record the 3 topic values $z = [z_1, z_2, z_3]$ and the 5 words $w = [w_1, w_2, w_3]$ generated. Then calculate the log probability of the document $p(w, z|\theta, \beta)$ given topic vector and word distributions. You need to show your derivation.