HW9 Recitation
Learning Paradigms

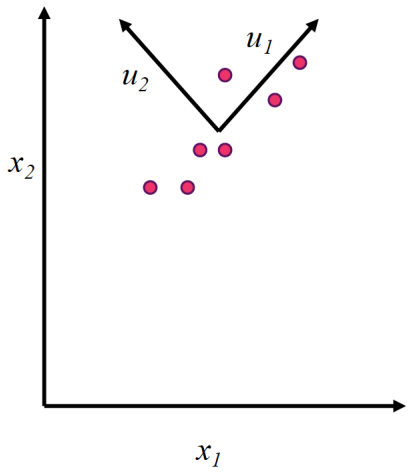
10-301/10-601: Introduction to Machine Learning

4/21/2023

1 Principal Component Analysis

Principal Component Analysis aims to project data into a lower dimension, while preserving as much as information as possible.

How do we do this? By finding an orthogonal basis (a new coordinate system) of the data, then pruning the "less important" dimensions such that the remaining dimensions minimize the squared error in reconstructing the original data.



In low dimensions, finding the principal components can be done visually as seen above, but in higher dimensions we need to approach the problem mathematically. We find orthogonal unit vectors $\mathbf{u}_1 \dots \mathbf{u}_M$ such that the reconstruction error $\frac{1}{N}\sum_{i=1}^{N} ||\mathbf{x}^{(i)} - \hat{\mathbf{x}}^{(i)}||^2$ is minimized, where $\hat{\mathbf{x}}^{(i)} = \sum_{m=1}^{M} (\mathbf{u}_m^T \mathbf{x}^{(i)})\mathbf{u}_m$ are the reconstructed vectors.

If we have M new vectors and d original vectors, with $M = d$, we can reconstruct the original data with 0 error. If $M < d$, it is usually not possible to reconstruct the original data without losing any error. In other words, all the reconstruction error comes from the $M - d$ missing components. This error can be expressed in terms of the covariance matrix of the original data, and is minimized when the principal component vectors $\mathbf{u}_1 \dots \mathbf{u}_M$ are the top M eigenvectors of the covariance matrix (in terms of eigenvalues). The higher the

eigenvalues for these eigenvectors are, the more information they store and the lower the reconstruction error.

For the following questions, use this Colab notebook.

Let's assume we've performed PCA on the following dataset:

Row	X1	X2	X3	X4
1	-0.21	-0.61	-0.35	0.08
2	0.15	-0.77	1.26	1.57
3	0.03	0.12	-0.39	-0.25
4	0.92	1.31	0.31	1.19
5	2.51	1.99	1.86	2.57
6	0.91	1.23	-0.01	0.04

And we've obtained the following principal components:

PC1	PC2	PC3	PC4
-0.53	0.23	0.48	-0.66
-0.49	0.7	-0.27	0.44
-0.43	-0.46	0.52	0.57
-0.54	-0.49	-0.65	-0.21

Which correspond to the following eigenvalues:

$$[3.265, 0.999, 0.043, 0.014]$$

1. Why are there only 4 principal components?

   There are 4 principal components because the original feature space has dimension 4. Thus, any new basis we construct can only have up to 4 independent components.

2. How much of the variance in the data is preserved by the first two principal components? (3.265 + 0.999) / (3.265 + 0.999 + 0.043 + 0.014) = 4.264 / 4.321 = 0.987 * 100 = 99% of the variance.

3. How much of the variance in the data is preserved by the first and third principal components? (3.265 + 0.043) / (3.265 + 0.999 + 0.043 + 0.014) = 3.308 / 4.321 = 0.766 * 100 = 76% of the variance.

4. Perform a dimensionality reduction on the points such that we project them onto the first two principal components. Then, inverse transform it back to four dimensions. What is the reconstruction error for this sample?

   The PCA'd dataset is:

$$\begin{bmatrix} 0.52 & -0.36 \\ -1.1 & -1.86 \\ 0.23 & 0.39 \\ -1.9 & 0.41 \\ -4.5 & -0.14 \\ -1.1 & 1.06 \end{bmatrix}$$

Projected back up to 4 dimensions, we get:

$$\begin{bmatrix} -0.36 & -0.5 & -0.06 & -0.1 \\ 0.16 & -0.77 & 1.33 & 1.51 \\ -0.03 & 0.16 & -0.28 & -0.32 \\ 1.1 & 1.21 & 0.64 & 0.83 \\ 2.36 & 2.09 & 2.02 & 2.5 \\ 0.83 & 1.28 & -0.01 & 0.07 \end{bmatrix}$$

Reconstruction error is 0.542.

5. Perform a dimensionality reduction such that we project the points onto the first and third principal components. Then, inverse transform it back to four dimensions. What is the reconstruction error of this new dataset?

   The new dataset is:

$$\begin{bmatrix} 0.52 & -0.17 \\ -1.1 & -0.08 \\ 0.23 & -0.06 \\ -1.9 & -0.52 \\ -4.5 & -0.03 \\ -1.1 & 0.07 \end{bmatrix}$$

   Projected back up to 4 dimensions, we get:

$$\begin{bmatrix} -0.36 & -0.21 & -0.32 & -0.17 \\ 0.54 & 0.56 & 0.43 & 0.65 \\ -0.15 & -0.1 & -0.13 & -0.09 \\ 0.76 & 1.07 & 0.55 & 1.37 \\ 2.37 & 2.2 & 1.94 & 2.45 \\ 0.62 & 0.52 & 0.52 & 0.54 \end{bmatrix}$$

   Reconstruction error is 5.259.

6. Consider the reconstruction error of the fourth row in particular. Is it lower using the first and second principal components or using the first and third? Why might this be the case?

   Using the first and second principal components:
   Error $= (0.92 - 1.1)^2 + (1.31 - 1.21)^2 + (0.31 - 0.64)^2 + (1.19 - 0.83)^2 = 0.28$

   Using the first and third principal components:
   Error $= (0.92 - 0.76)^2 + (1.31 - 1.07)^2 + (0.31 - 0.55)^2 + (1.19 - 1.37)^2 = 0.17$

   This is because PCA minimizes the mean reconstruction error over all rows, so there may be rows/data points whose reconstruction errors are not minimized (i.e. another choice of projection might yield lower error for those points).

2 K-Means

Clustering is an example of unsupervised machine learning algorithm because it serves to partition **unlabeled** data. There are many different types of clustering algorithms, but the one that is used most frequently and was introduced in class is **K-Means**.

In K-Means, we aim to minimize the objective function:

$$\sum_{i=1}^{n} \min_{j \in \{1,...,k\}} ||\mathbf{x}^{(i)} - \mathbf{c}_j||^2 \tag{1}$$

Below is the K-Means algorithm:

Let $\mathcal{D} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(n)}\}$ where $\mathbf{x}^{(i)} \in \mathbb{R}^d$ be the set of input examples that each have d features.

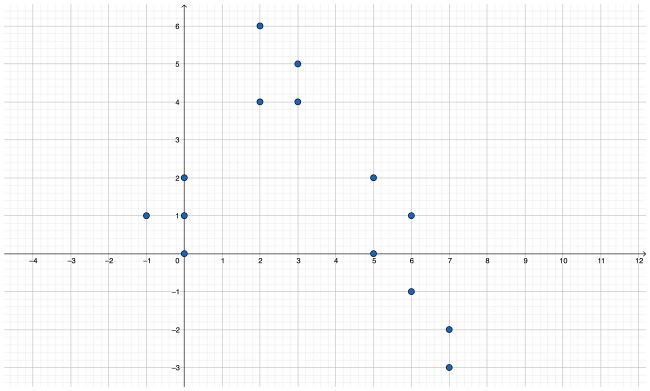
Initialize k cluster centers $\{\mathbf{c}^{(1)}, ..., \mathbf{c}^{(k)}\}$ where $\mathbf{c}^{(i)} \in \mathbb{R}^d$

Repeat until convergence:

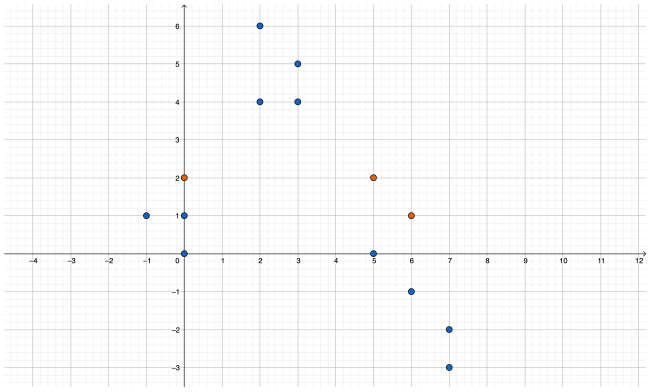
1. Assign each point $\mathbf{x}^{(i)}$ to a cluster $\mathcal{C}^{(j)}$ where $j = \arg\min_{1 \leq r \leq k} ||\mathbf{x}^{(i)} - \mathbf{c}^{(r)}||$
2. Recompute each $\mathbf{c}^{(i)}$ as the mean of points in $\mathcal{C}^{(i)}$

2.1 Walking through an example

Lets walk through an example of K-Means with $k = 3$ using the following dataset for the first iteration:



Let the cluster centers be initialized to $\mathbf{c}^{(1)} = (0, 2)$, $\mathbf{c}^{(2)} = (5, 2)$, $\mathbf{c}^{(3)} = (6, 1)$ as depicted below in the orange:



Perform one iteration of the K-Means algorithm:

1. What are the cluster assignments? $\mathcal{C}^{(1)} = \{(0,0), (-1,1), (0,1), (0,2), (2,4), (2,6)\}$

   $\mathcal{C}^{(2)} = \{(3,4), (3,5), (5,2)\}$

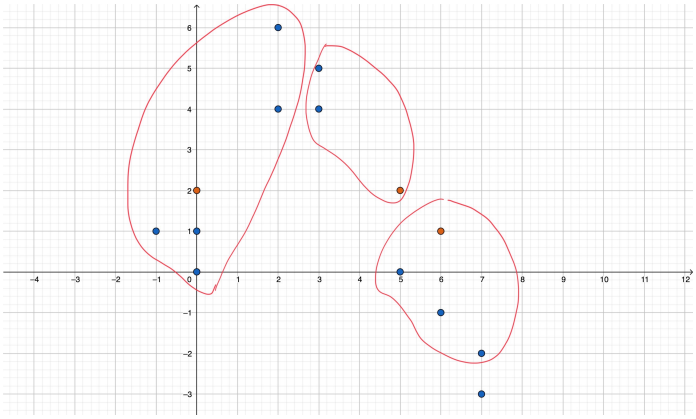
   $\mathcal{C}^{(3)} = \{(5,0), (6,1), (6,-1), (7,-2), (7,-3)\}$

2. What are the recomputed cluster centers? $\mathbf{c}^{(1)} = (0.5, 2.33)$

   $\mathbf{c}^{(2)} = (3.67, 3.67)$

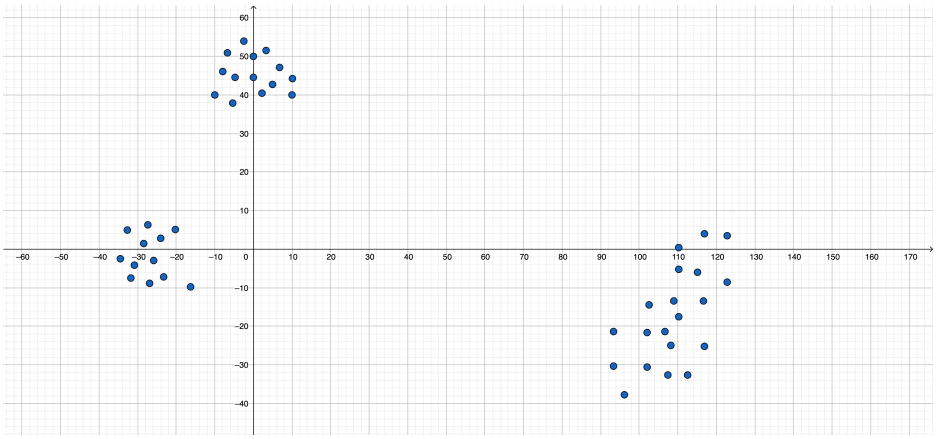
   $\mathbf{c}^{(3)} = (6.2, -1)$

3. Draw the cluster assignments after the first iteration on the graph below.



2.2 The importance of initialization

Given the points in the graph below, and assume we will have $k = 3$ cluster centers.



1. Given an example of a set of initialization points such that the K-Means algorithm would converge to a global minimum.

   Any three points where each belongs to a different cluster

2. Given an example of a set of initialization points such that the K-Means algorithm would converge to a local minimum instead of the global minimum.

   For example, one to the upper left corner, the other two at the bottom right corner

3 Ensemble Methods

The idea of ensemble methods is to build a model for prediction by combining the strengths of a group of simpler models. We'll cover two examples of ensemble methods: random forests and AdaBoost.

3.1 Random Forests

1. What are some downsides of decision trees, and how can we explain this in the context of the bias-variance tradeoff?

   learned greedily, can overfit if depth isn't controlled, low bias but high variance

$$\text{Random Forests} = \text{Sample Bagging} + \text{Split-Feature Randomization}$$

2. What is **sample bagging**?

   Bagging stands for bootstrap aggregating. A bootstrapped dataset has the same number of rows as the original dataset, but its rows are drawn from the original dataset with replacement. Models are trained on each individual dataset, but since the training datasets are not as similar to each other, the learned models tend to be different and more variable as well. Aggregating refers to the model predictions being combined to form the final prediction.

3. What is **the random subspace method**?

   In the splits for each decision tree, instead of choosing the split feature from the set of all features, we limit the feature set to a randomly chosen subset of all the features. This further reduces correlation between the learned decision trees as the "best" feature may not always be available to split on.

4. How do these techniques affect the bias and variance of an individual tree?

   Both increase bias and decrease variance by limiting the information available to train on, compared to a decision tree trained on the full original dataset.

5. How do these techniques affect the bias and variance of an ensemble of trees?

   The increase in bias carries over from the individual trees. The ensemble has lower variance because we are aggregating predictions over a set of trees that are not fully correlated. Both techniques are designed to make the trees less correlated with one another, as reducing covariance between trees reduces variance of the average over trees.

Consider random variables X_1, X_2, each with variance σ^2. The variance of their average is given by

$$\begin{aligned}
\text{Var}(\frac{1}{2}X_1 + \frac{1}{2}X_2) &= \text{Var}(\frac{1}{2}X_1) + \text{Var}(\frac{1}{2}X_2) + 2\text{Cov}(\frac{1}{2}X_1, \frac{1}{2}X_2) \\
&= \frac{1}{4}\text{Var}(X_1) + \frac{1}{4}\text{Var}(X_2) + \frac{1}{2}\text{Cov}(X_1, X_2) \\
&= \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 + \frac{1}{2}\sigma^2\frac{\text{Cov}(X_1, X_2)}{\sigma \cdot \sigma} \\
&= \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 + \frac{1}{2}\rho\sigma^2 \\
&= \sigma^2\frac{1}{2}(1+\rho)
\end{aligned}$$

where ρ is the correlation between X_1, X_2.

If $\rho = 1$ (fully correlated trees), then we achieve no variance reduction: the average has variance σ^2. In general, reducing ρ makes the trees less correlated and reduces variance.

Note that our techniques are generally not extreme enough to generate anticorrelated trees, where tree predictions would oppose each other. The more anticorrelated our predictions are, the closer our model gets to always predicting 0, which does reduce variance, but at the cost of performance on our task.

6. For each data point $\mathbf{x}^{(i)}$, define $t^{(-i)}$ to be the set of decision trees that $\mathbf{x}^{(i)}$ was not used to train. Use each tree in $t^{(-i)}$ to make a prediction for $\mathbf{x}^{(i)}$, and use these predictions to make an aggregated prediction $\overline{t^{(-i)}}(\mathbf{x}^{(i)})$ (i.e. for classification take the majority vote). Then, we can define the *out-of-bag* error as follows:

$$E_{OOB} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(\overline{t^{(-i)}}(\mathbf{x}^{(i)}) \neq y^{(i)}\right)$$

Why can we use E_{OOB} for hyperparameter optimization even though it was calculated using training points we used to learn the decision trees with?

While every point was used to train a certain set of decision trees, the calculation of E_{OOB} takes advantage of the fact that the nature of bootstrapped datasets means that there will generally be a reasonably large proportion of them that do not contain any particular point.

Therefore, for the decision trees that were trained on these datasets, the training point is equivalent to a test point as the tree has never seen it before. Since each tree is trained independently, there will never be a scenario in which a tree is both trained on a data point and also evaluated on it.

7. **Random Forest Example:** Suppose we train a random forest with two decision trees on the following dataset, using the provided bootstrap samples. Assume that for ties, we predict $Y = 1$.

All	X_0	X_1	X_2	X_3	Y
1	1	0	0	0	1
2	0	0	1	0	1
3	0	0	0	1	1
4	0	0	0	0	0
5	0	1	0	1	1

Sample 1	X_0	X_1	X_2	X_3	Y
1	1	0	0	0	1
4	0	0	0	0	0
5	0	1	0	1	1

Sample 2	X_0	X_1	X_2	X_3	Y
3	0	0	0	1	1
4	0	0	0	0	0
5	0	1	0	1	1

(a) Suppose we train our first tree on Sample 1 and the split feature randomization chooses $\{X_1, X_2\}$ for the feature candidates at the root. What feature will we split on at the root? X_1

(b) Suppose we then recurse on the left child (with feature value 0) of the root and split feature randomization chooses $\{X_0, X_2\}$ for the feature indices. What feature will we split on? X_0

(c) Suppose we train our second tree on Sample 2 and the split feature randomization chooses $\{X_2, X_3\}$ for the feature candidates at the root. What feature will we split on at the root? X_3

(d) What is the training error of the ensemble? 1/5, as only point 2 is incorrect.

Point 1: tree 1 predicts 1, tree 2 predicts 0, so prediction is 1

Point 2: tree 1 predicts 0, tree 2 predicts 0, so prediction is 0

Point 3: tree 1 predicts 0, tree 2 predicts 1, so prediction is 1

Point 4: tree 1 predicts 0, tree 2 predicts 0, so prediction is 0

Point 5: tree 1 predicts 1, tree 2 predicts 1, so prediction is 1

(e) What is the out of bag error of the ensemble? 4/5, as only point 5 is correct

Point 1: only tree 2 is involved, prediction is 0

Point 2: both trees are involved, prediction is 0

Point 3: only tree 1 is involved, prediction is 0

Points 4 and 5: majority vote over 0 trees predicts 1

3.2 AdaBoost

3.2.1 AdaBoost Weighting

AdaBoost relies on building an ensemble of weak learners, assigning them weights based on their errors during training.

1. Assume we are in the binary classification setting. What happens to the weight $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$ of classifier h_t if its error $\epsilon_t > 0.5$? Why is this useful? It becomes negative (check the log term). This "inverts" the output of this weak learner for every input, turning it into a classifier with $\epsilon_t < 0.5$.

   Note that if we can find weak learners h_t with $\epsilon_t < 0.5$ for all t, training error will decrease exponentially fast in the total number of iterations T.

2. AdaBoost also assigns weights $\mathcal{D}_t(i)$ for each data point. Explain in broad terms how the weights assigned to examples get updated in each iteration.

   Generally, points that get incorrectly classified get up-weighted and points that get correctly classified get down-weighted. The amount by which they're weighted depends on the importance of the weak learner - better (lower error) learners lead to stronger up-weights and down-weights. The weights are also normalized to have sum 1.

   Update rule: $\mathcal{D}_t(i) \propto \mathcal{D}_{t-1}(i)\exp\left(-\alpha_t y^{(i)} h_t(x^{(i)})\right)$

3.2.2 Weak Learners

We always talk using AdaBoost with "weak" learners; why can't we ensemble together "stronger" learners? Let's take a look at bounds on the test error of AdaBoost, fixing the number of samples N and number of training iterations T, but allowing variation in the hypothesis class of learners $\mathcal{H}$.

Let d be the VC-dimension of the hypothesis class. Consider the following bounds with respect to d:

$$\text{Bound 1 (PAC Learning)}: \quad \epsilon(H_T) \leq \hat{\epsilon}_S(H_T) + O\left(\sqrt{T \log T}\sqrt{d}\sqrt{\frac{\log N}{N}}\right)$$

$$\text{Bound 2 (Margin Analysis)}: \quad \epsilon(H_T) \leq \hat{P}_S\left[\text{margin}_T \leq \theta\right] + O\left(\frac{1}{\theta}\sqrt{d}\sqrt{\frac{\log^2 N}{N}}\right)$$

1. What happens to our bounds on true error if we increase the VC dimension of the weak learner hypothesis space? The bounds loosen/increase.
2. What concept does this connection between classifier complexity and error relate to? Overfitting

4 Recommender Systems

4.1 Collaborative Filtering

Collaborative filtering recommends items to users based on other similar users' preferences, meaning that it depends on the ratings to an item from other users. We have covered two types of collaborative filtering methods in the lecture:

- Neighborhood Methods
- Latent Factor Methods (e.g., Matrix Factorization)

4.1.1 Neighborhood Methods

Neighborhood methods in collaborative filtering extract a neighborhood given the user data (the items you have experienced) and recommend the items preferred by this neighborhood to the user. The step-by-step approach is:

1. Observe the items the target user has experienced
2. Find the other user or users who have experienced the most of those items
3. Recommend the set of items not experienced by the target user that have been experienced by the largest number of these other users

Let's assume for each user, we can construct a following vector:

$$U_{items} = \{u_1, u_2, \cdots, u_k\}, u_i = \begin{cases} 1 & \text{if the user has viewed item } i \\ 0 & \text{if the user has not viewed item } i \end{cases}$$

Is the closest neighbor by Manhattan or Euclidean distance of U_{items} vectors always in the neighborhood we use for recommendations?

No. Because we want our closest neighbor of our target user to be chosen only based on their shared experiences with our target user. We also want these neighbors to have experiences which the target user does not have, so that we can recommend unseen items. Thus, we do not want to penalize by increasing distance if the other users have experienced many things the target user has not.

4.1.2 Matrix Factorization

1. When doing PCA, given a dataset X, we are able to perform SVD to find the eigenvectors and eigenvalues of the covariance matrix $\frac{1}{N}X^TX$. If dealing with a user/item matrix $R \in \mathbb{R}^{n\times m}$, can we also use SVD to find the matrix decomposition R into user matrix $U \in \mathbb{R}^{n\times d}$ and item matrix $V \in \mathbb{R}^{m\times d}$? If yes, write out the formula for the decomposition; if no, explain why not.

   We cannot use SVD to find the decomposition of the user/rating matrix, because the R matrix has missing values. SVD cannot be performed on a dataset with missing values.

4.1.3 Alternating Least Squares for Matrix Factorization

Because both U and V are unknowns, our objective function is non-convex and hard to optimize. However, if we fix one of the unknowns, the optimization problem becomes quadratic and can be directly solved. ALS rotates between fixing the U to optimize V and fixing V to optimize U. This algorithm is called **Block Coordinate Descent**.

1. If we fix one of the unknowns, what known problem (with a closed-form solution) does this reduce to?

   Least squares (from linear regression), hence the name.

2. Write the block coordinate descent pseudocode for ALS.

$$\mathbf{U} \leftarrow \arg\min_{\mathbf{U}} J(\mathbf{U}, \mathbf{V})$$
$$\mathbf{V} \leftarrow \arg\min_{\mathbf{V}} J(\mathbf{U}, \mathbf{V})$$

   Now, let's look at the interpretation of our user and item vectors. Note that both types of vector inhabit the shared coordinate space $\mathbb{R}^d$, and that we compute similarity with a dot product. This allows us to interpret both user vectors and item vectors as representations in a shared lower-dimensional space.

4.2 Content-Based Filtering

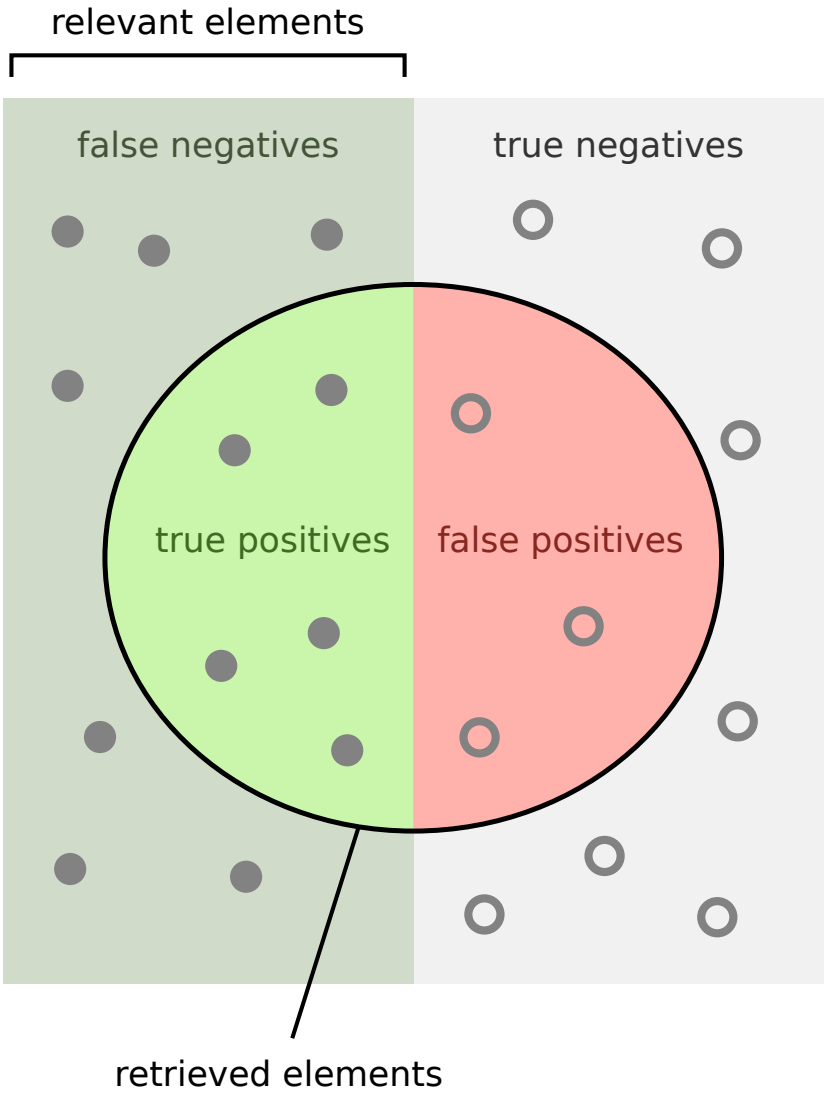
1. Suppose we are trying to recommend movies to a user. We are given a feature vector for each movie with content information such as year of release and genre, and for movies the user has watched, we are given labels for whether or not they liked the movie. What learning paradigm is suited for our recommendation task? Supervised learning. Train a model on the features and labels and make predictions on unseen movies.

2. What is one advantage of content-based filtering over collaborative filtering? We don't need other users in the system at all; can start making predictions without user/item interactions. We also don't need to take and store user data, which improves data

privacy. This can also be much more explainable than specifically MF for collaborative filtering, since we can choose our features instead of having the optimally computed but hard to understand lower-dimensional space.

3. What is one advantage of collaborative filtering over content-based filtering? It can be hard or computationally expensive to find or compute content information.

5 Precision and Recall



How many retrieved items are relevant?

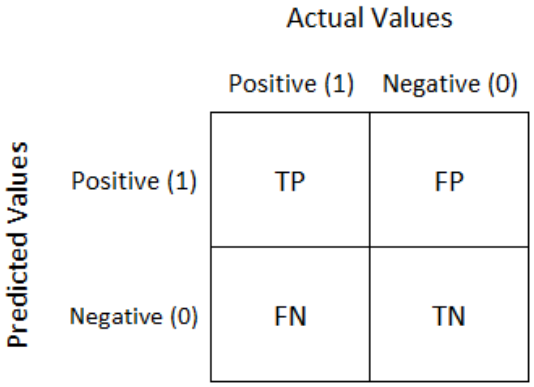
$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are retrieved?

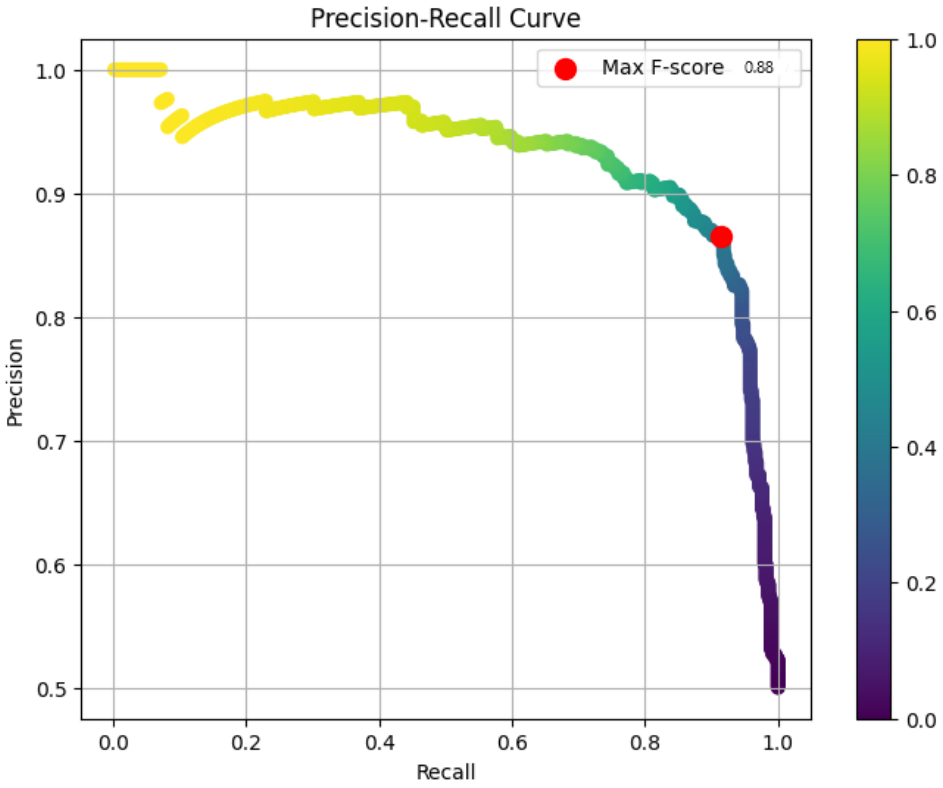
$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

The following chart is known as a *confusion matrix* and helps formalize the concepts displayed above. There are 4 categories in the chart:

- *true positives*: items that are predicted positive and have actual label positive
- *false positives*: items that are predicted positive but have actual label negative
- *true negatives*: items that are predicted negative and have actual label negative
- *false negatives*: items that are predicted negative but have actual label positive

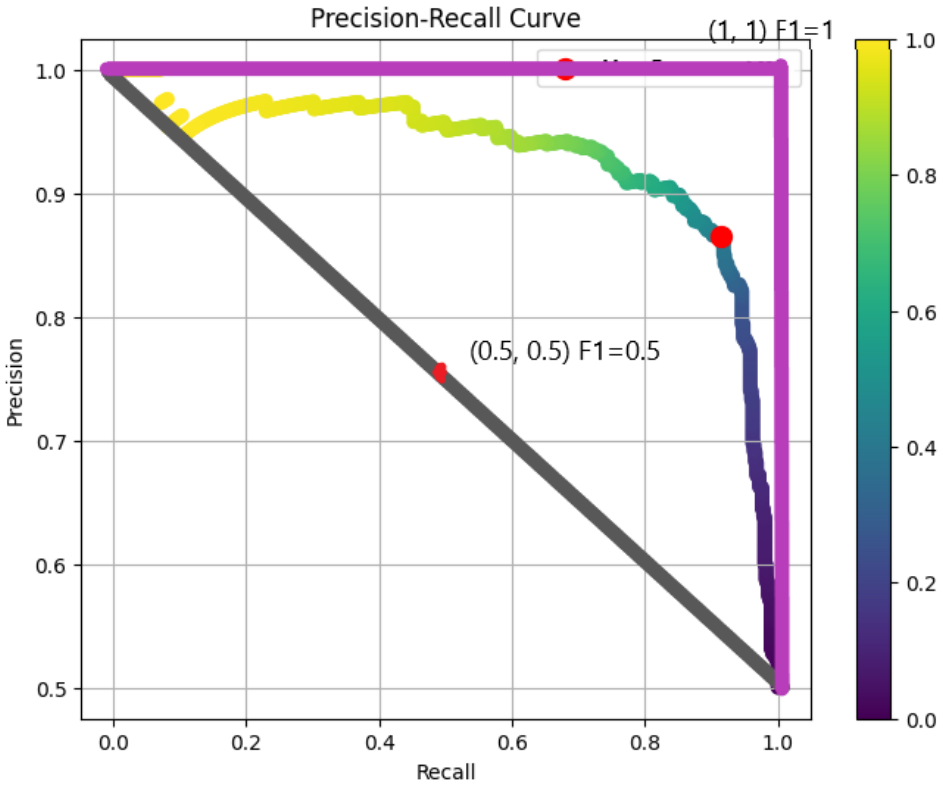


1. What is the formula for precision in terms of the values in the confusion matrix? What about recall? Precision = TP/(TP + FP), Recall = TP/(TP + FN)
2. The *base rate* is the proportion of items that have true label positive. What is the formula for the base rate in terms of the confusion matrix? base rate = (TP + FN) / (TP + FP + FN + TN)
3. Suppose we predict every item to be positive. What is the precision? What is the recall? precision = base rate, recall = 1
4. The F_1 score is defined as the harmonic mean of the precision and recall: $F_1 = \frac{2}{1/P+1/R}$. The following image shows an example curve of precision and recall for a classifier when varying the threshold between the positive and negative classes. The point on the curve with highest F_1 score is marked.



Draw an example precision-recall curve for a "better" classifier than the one shown. Mark the point with the optimal F_1 score.

Draw an example precision-recall curve for a "worse" classifier than the one shown. Mark the point with the optimal F_1 score.


Precision-Recall Curve
(1, 1) F1=1
(0.5, 0.5) F1=0.5
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.2
0.4
0.6
0.8
1.0