Classification

T.L. Grobler

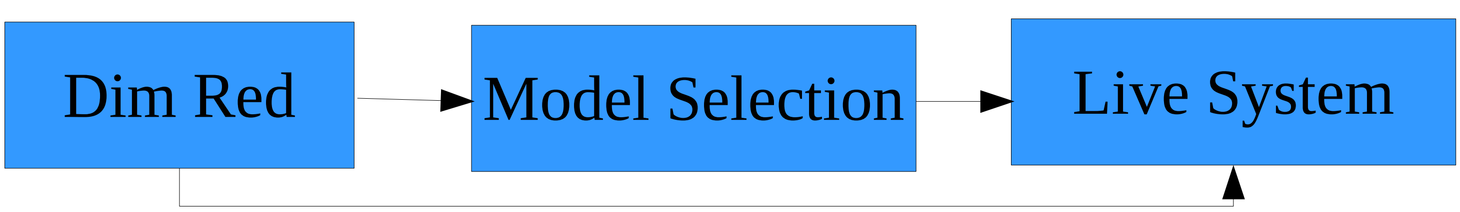
Outline

- Introduction to supervised classifiation (labelled data):
  - Generative: Naive-Bayes
  - Discriminative: Logistic Regression
- Introduction to unsupervised classification:
  - *k*-means
  - GMM

Data Sets & Preprocessing

- **Dimensionality reduction:**
  - PCA
  - LDA
- **Data Sets:**
  - **Training Set:** Train Classifier
  - **Validation Set:** Hypeparameters (prevent overfitting).
  - **Test Set:** Performance.

NB: Be careful-
Splitting of data



Performance

- Supervised:
  - confusion matrix
  - Accuracy
  - etc...
- Unsupervised:
  - GMM: likelihood
  - K-means: distortion

Supervised Classification (Chap 4)

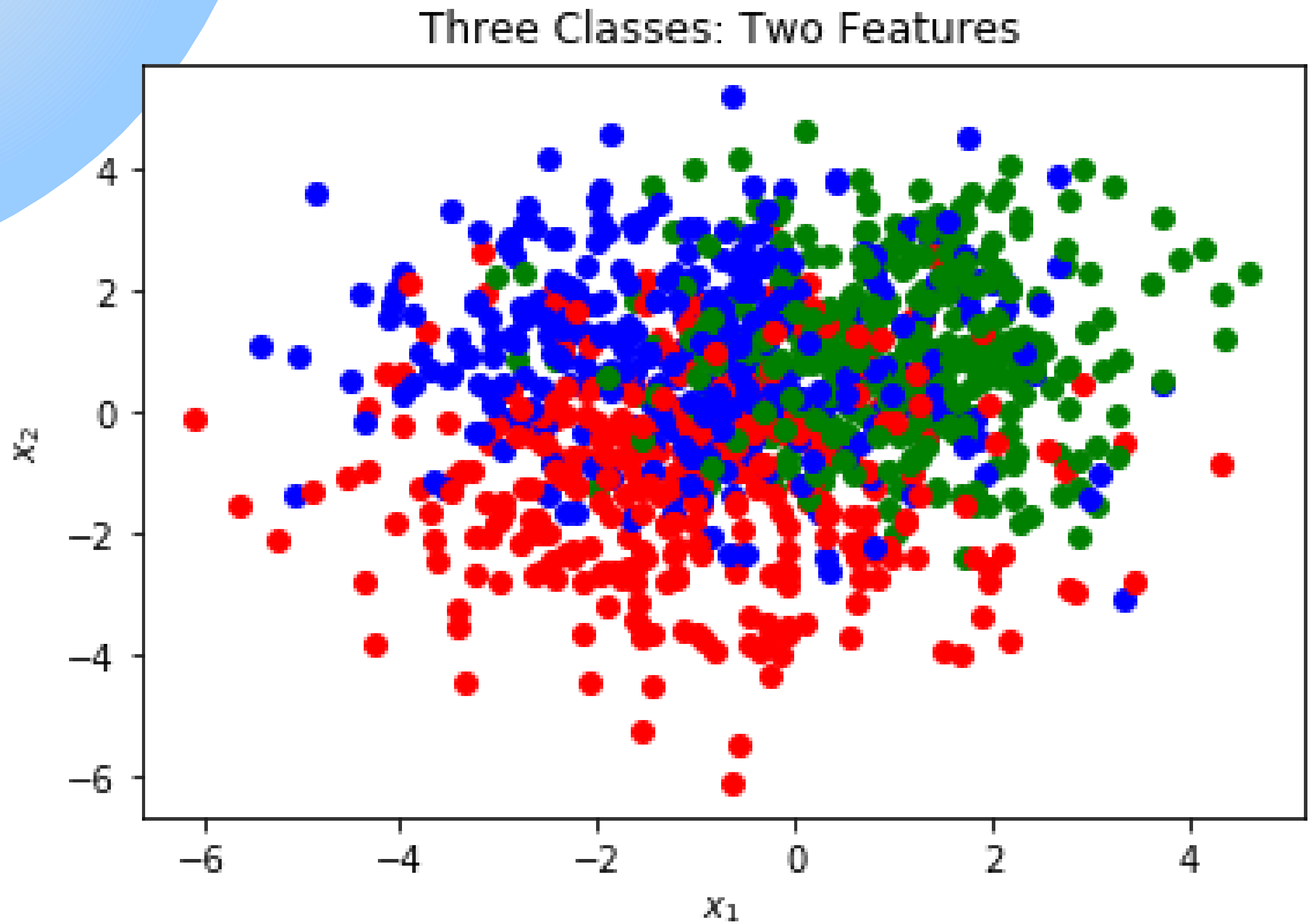
Classification/Classifiers

- Given $\mathbf{x}$ assign to one of k classes:
  - C_j, $j = 1,\ldots,k$
  - Assign prob $P(C_j|\mathbf{x})$
  - $C^* = \mathrm{argmax}_{C_j}\, P(C_j|\mathbf{x})$
- Class prob, more useful than knowing max class prob.

Data Description

- D: $(\mathbf{x}_j, y_j)$, $j = 1,\ldots,N$
  - observation $\mathbf{x}_j$ comes with class label y_j
  - $y_j = C_j$ if $\mathbf{x}_j \in C_j$
- Constructing $P(C_j|\mathbf{x})$ given D
- Two approaches: *generative* and *discriminative*

Example Training Data



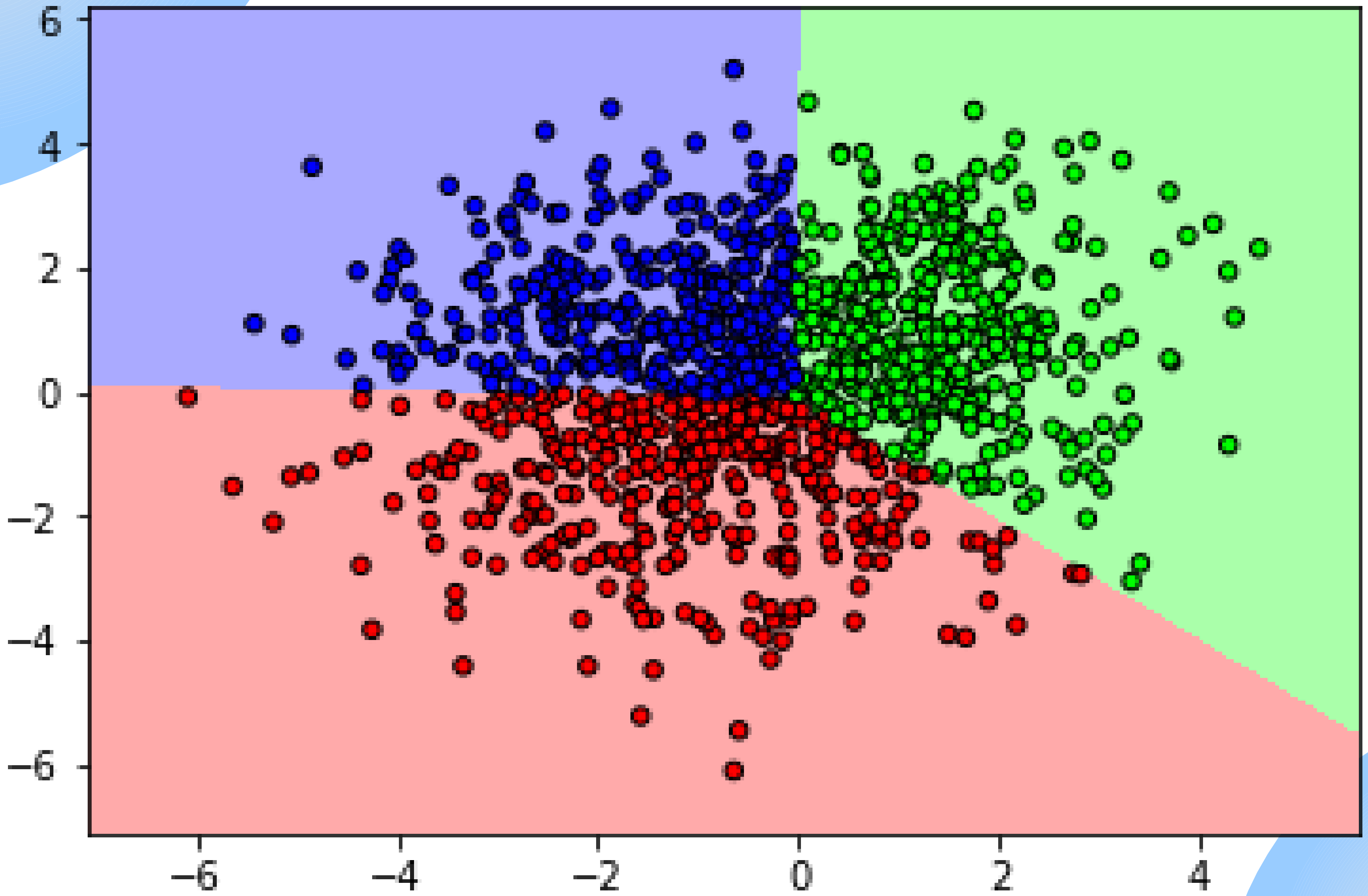
$\mathbf{x} = [x_1, x_2]$ - features

C_1 = r, C_2 = g and C_3 = b

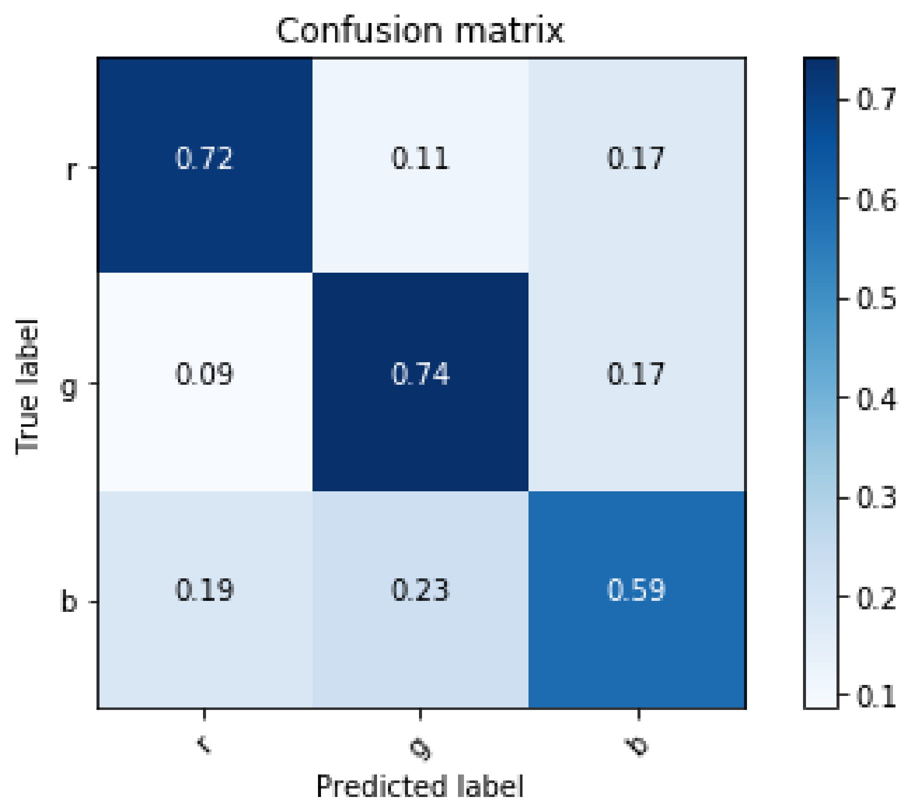
Nearest Centroid Classifier

- Imports:
  - `from sklearn.neighbors.nearest_centroid import NearestCentroid`
  - `from sklearn.metrics import confusion_matrix`
- Train:
  - `clf = NearestCentroid()`
  - `clf.fit(X, y)`
- Predict:
  - `y_pred = clf.predict(X)`
- Confusion Matrix:
  - `cm = confusion_matrix(y,y_pred)`

Decision Boundary



Confusion Matrix



M_{ij} is % of observations in class i predicted to be in class j.

Generative Approach

- Estimate class-conditionals: $p(\mathbf{x}|C_j)$
- Posterior (Bayes Theorem):
  - $P(C_j|\mathbf{x}) \propto p(\mathbf{x}|C_j)P(C_j)$
  - Introduce $P(C_j)$
- Probabilistic Generative Models (PGM)

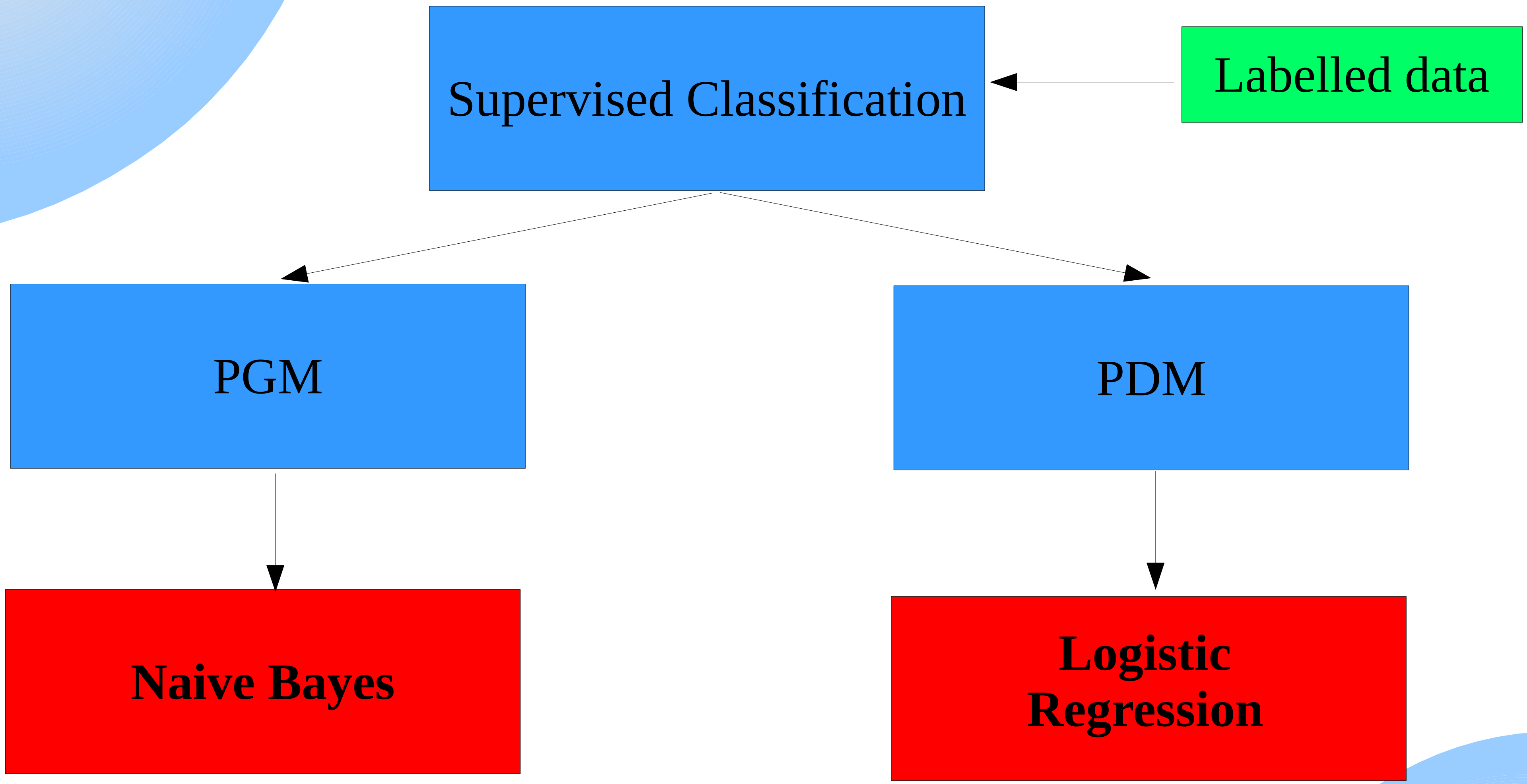
Discriminative Approach

- Dispenses with: $p(\mathbf{x}|C_j)$
- Directly compute posterior $P(C_j|\mathbf{x})$
- Probabilistic Discriminitive Models (PDM)

Generative vs Discriminitive

Generative	Discriminitive
Wasteful	Unwasteful
Training Straightforward	Training Harder
Ignores Class Dependence	Incorporates Class Dependence
Class Data	All Data
Separate Classes	Max Class Differences

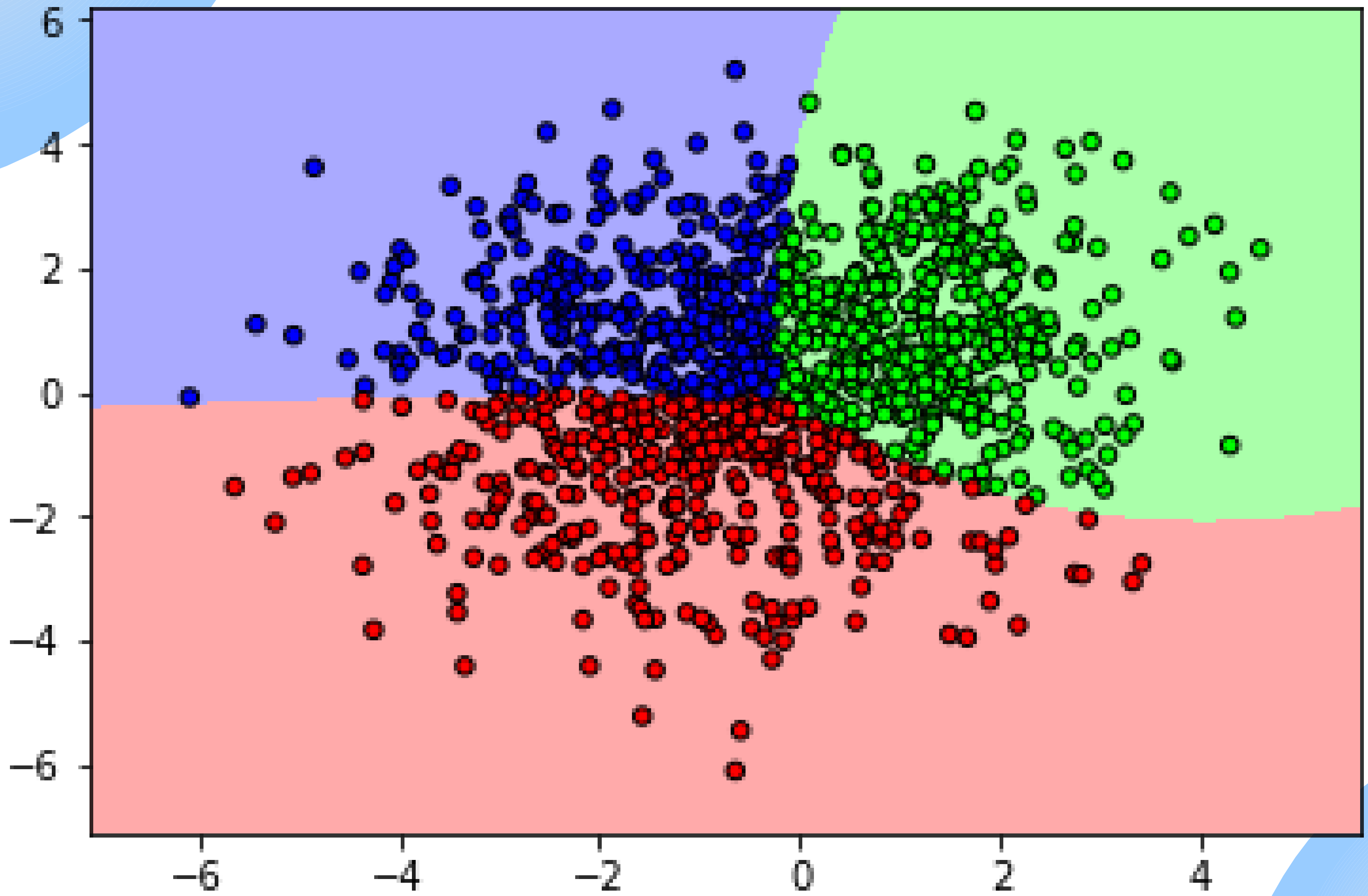
Supervised Classification



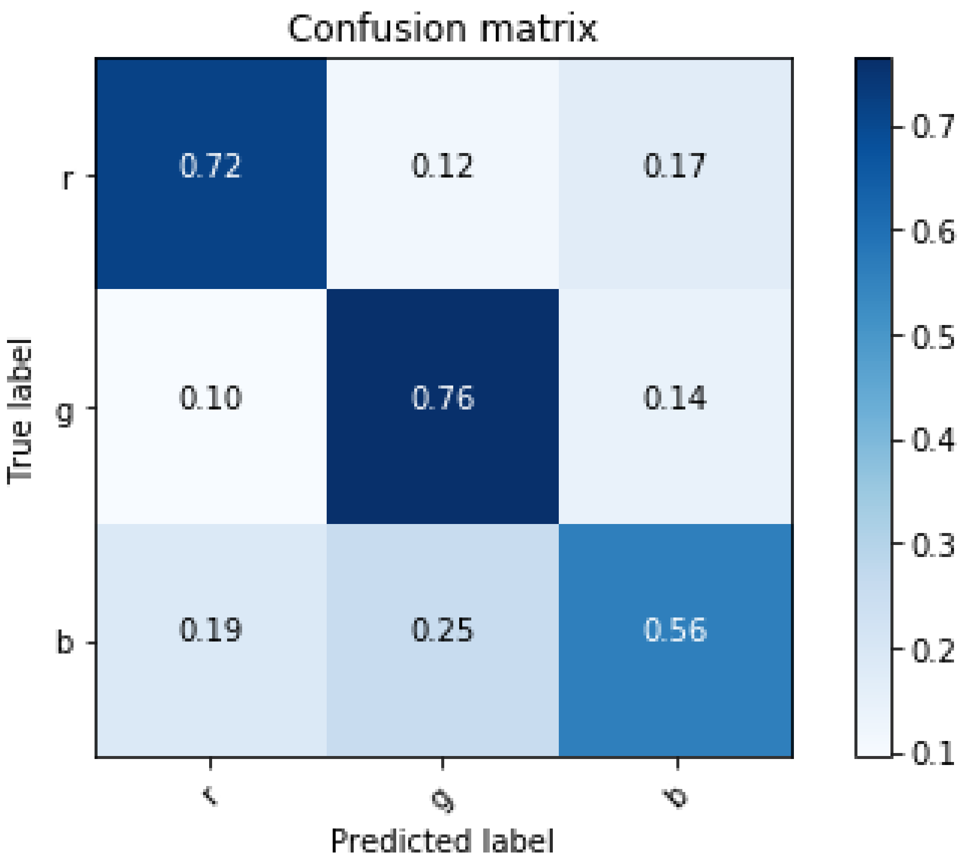
Naive Bayes

- Imports:
  - `from sklearn.naive_bayes import GaussianNB`
  - `from sklearn.metrics import confusion_matrix`
- Train:
  - `clf = GaussianNB()`
  - `clf.fit(X, y)`
- Predict:
  - `y_pred = clf.predict(X)`
- Confusion Matrix:
  - `cm = confusion_matrix(y,y_pred)`

Decision Boundary



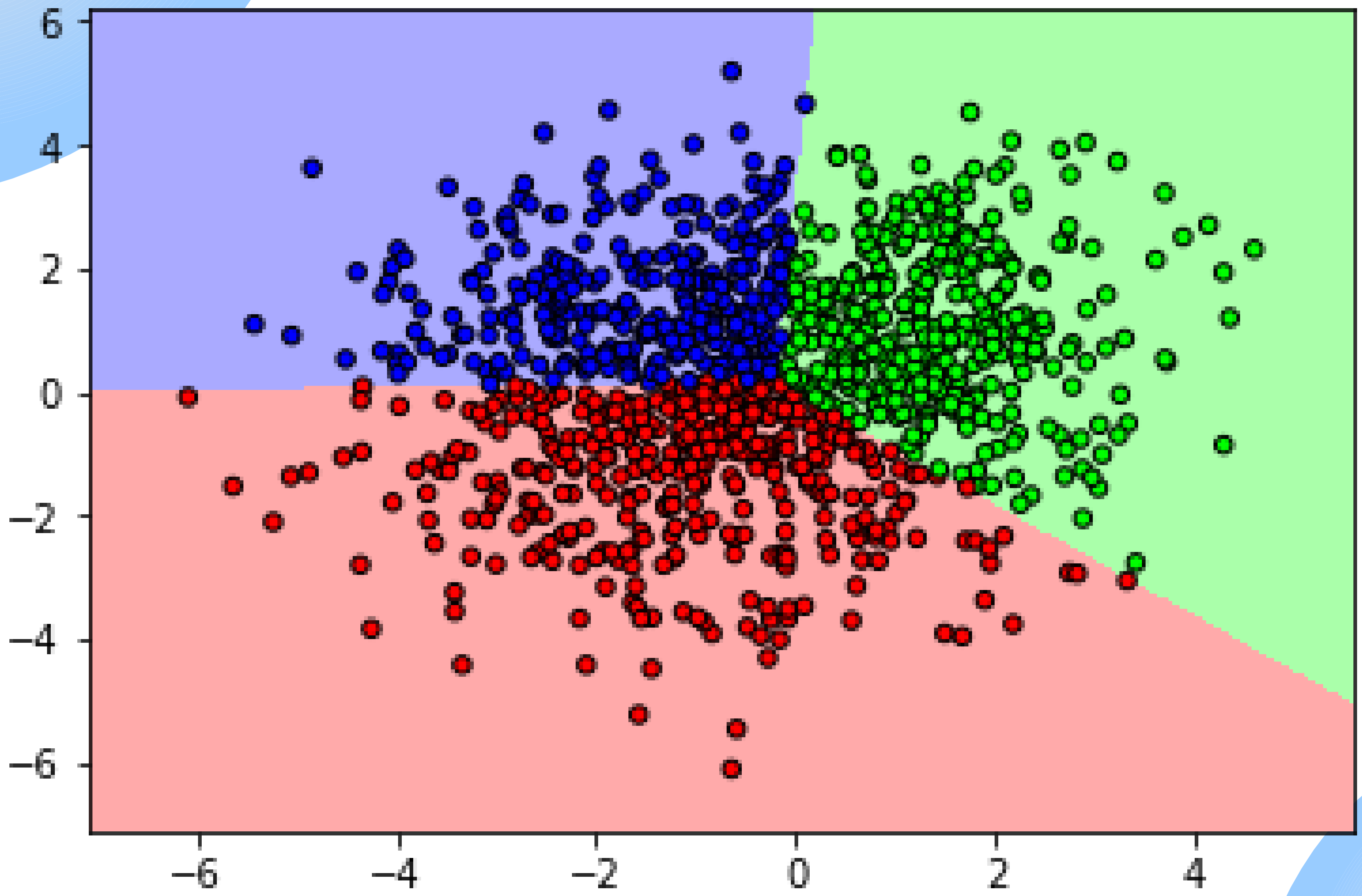
Confusion Matrix



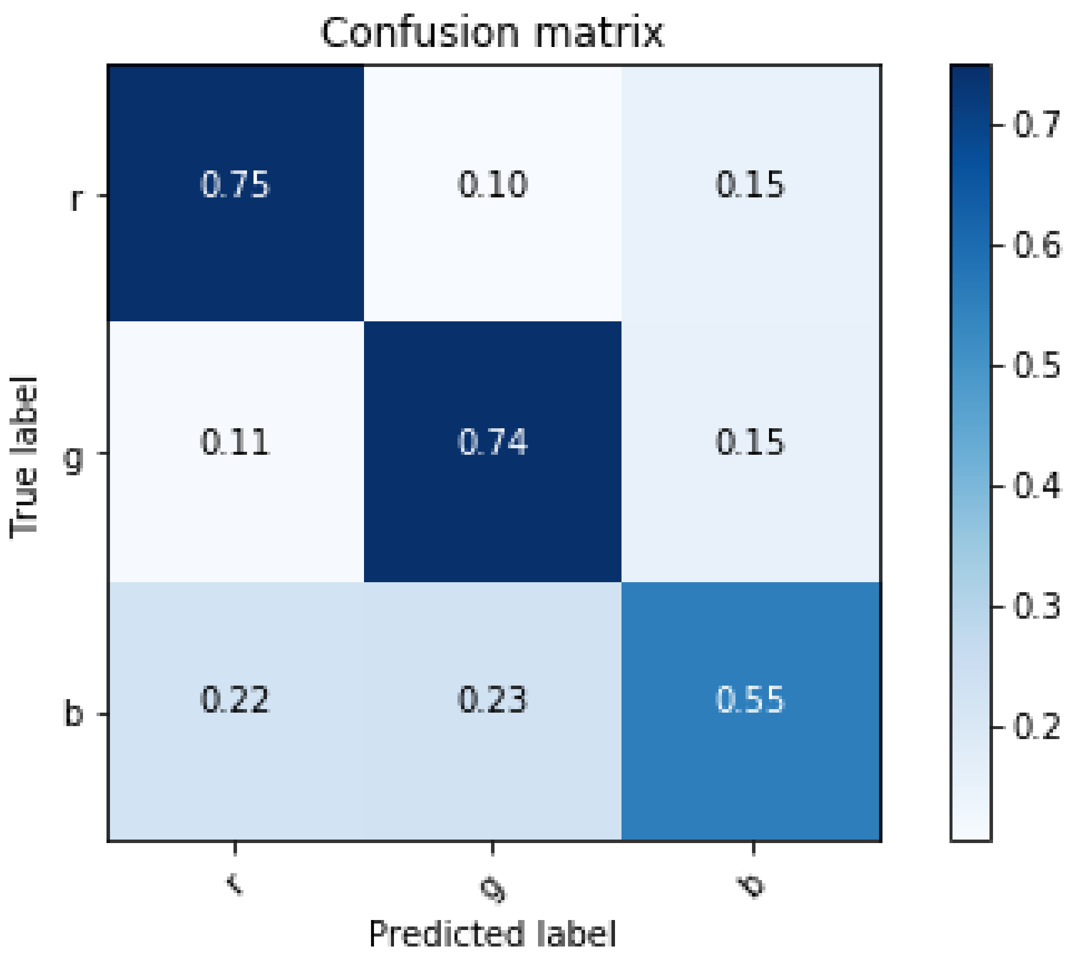
Logistic Regression

- Imports:
  - `from sklearn.linear_model import LogisticRegression as logis`
  - `from sklearn.metrics import confusion_matrix`
- Train:
  - `clf = linear_model.LogisticRegression(C=1e5)`
  - `clf.fit(X, y)`
- Predict:
  - `y_pred = clf.predict(X)`
- Confusion Matrix:
  - `cm = confusion_matrix(y,y_pred)`

Decision Boundary

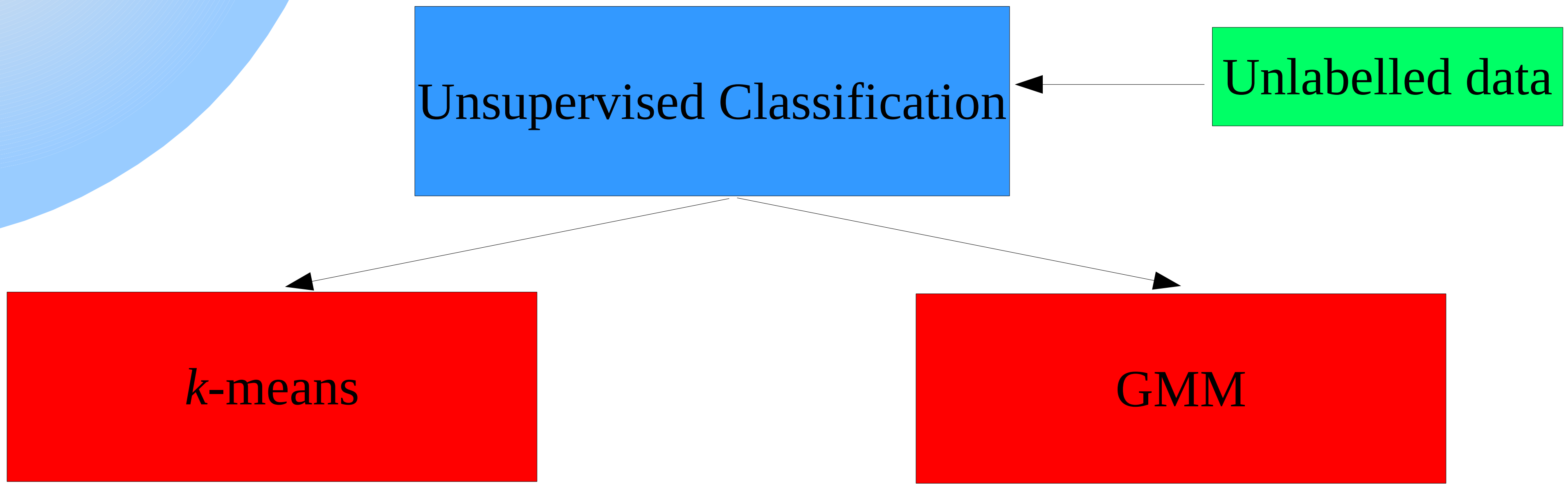


Confusion Matrix



Unsupervised Classification (Chap 5)

Unsupervised Classification



- Clustering data into *k-clusters* without any class labels. Unlabelled data.

k-means

- Imports:
  - `from sklearn.cluster import KMeans`
  - `from sklearn.metrics import confusion_matrix`
- Train:
  - `kmeans = KMean(n_clusters=3)`
  - `kmeans.fit(X)`

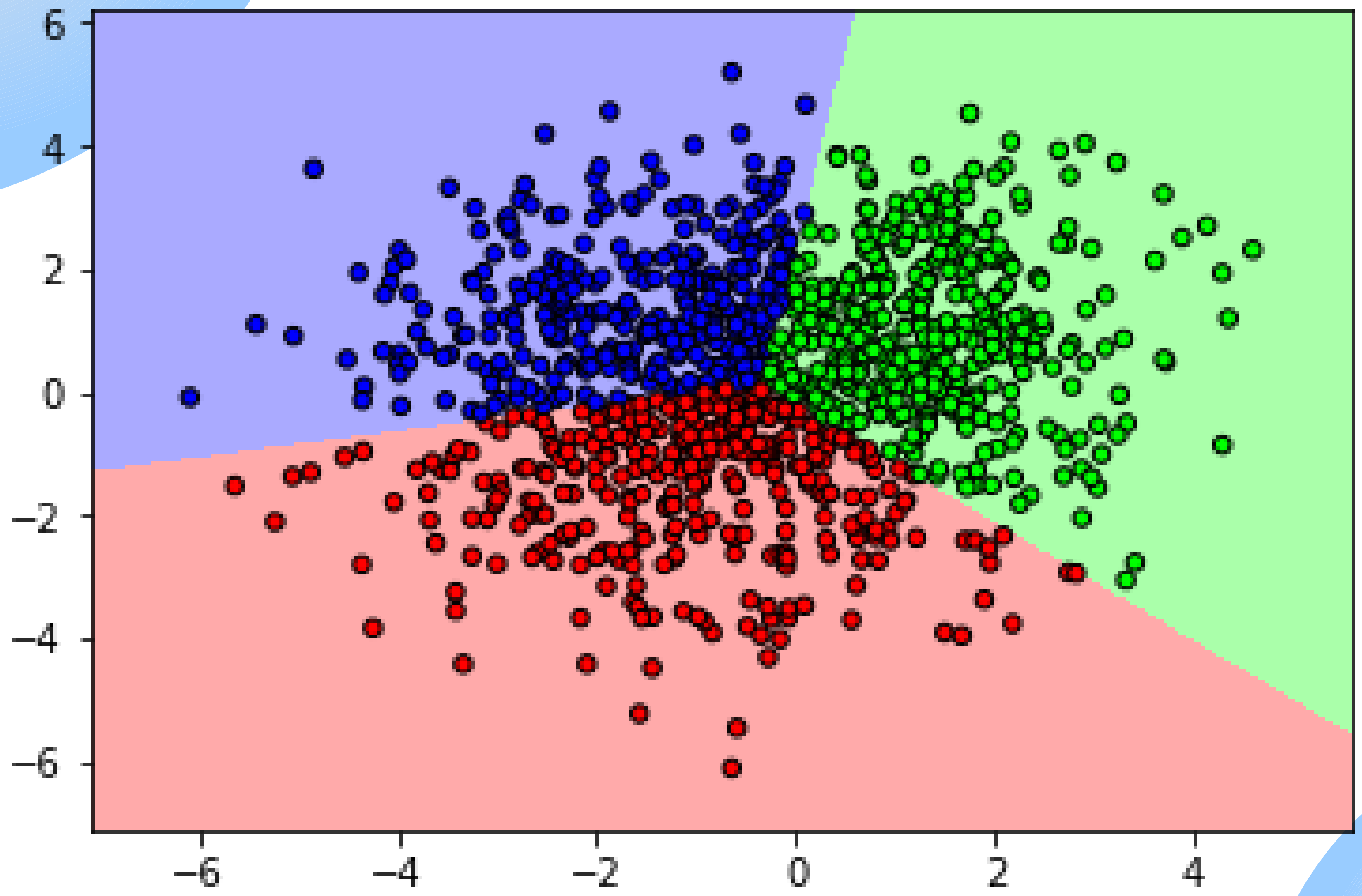
No-labels

- Predict:
  - `y_pred = kmeans.predict(X)`

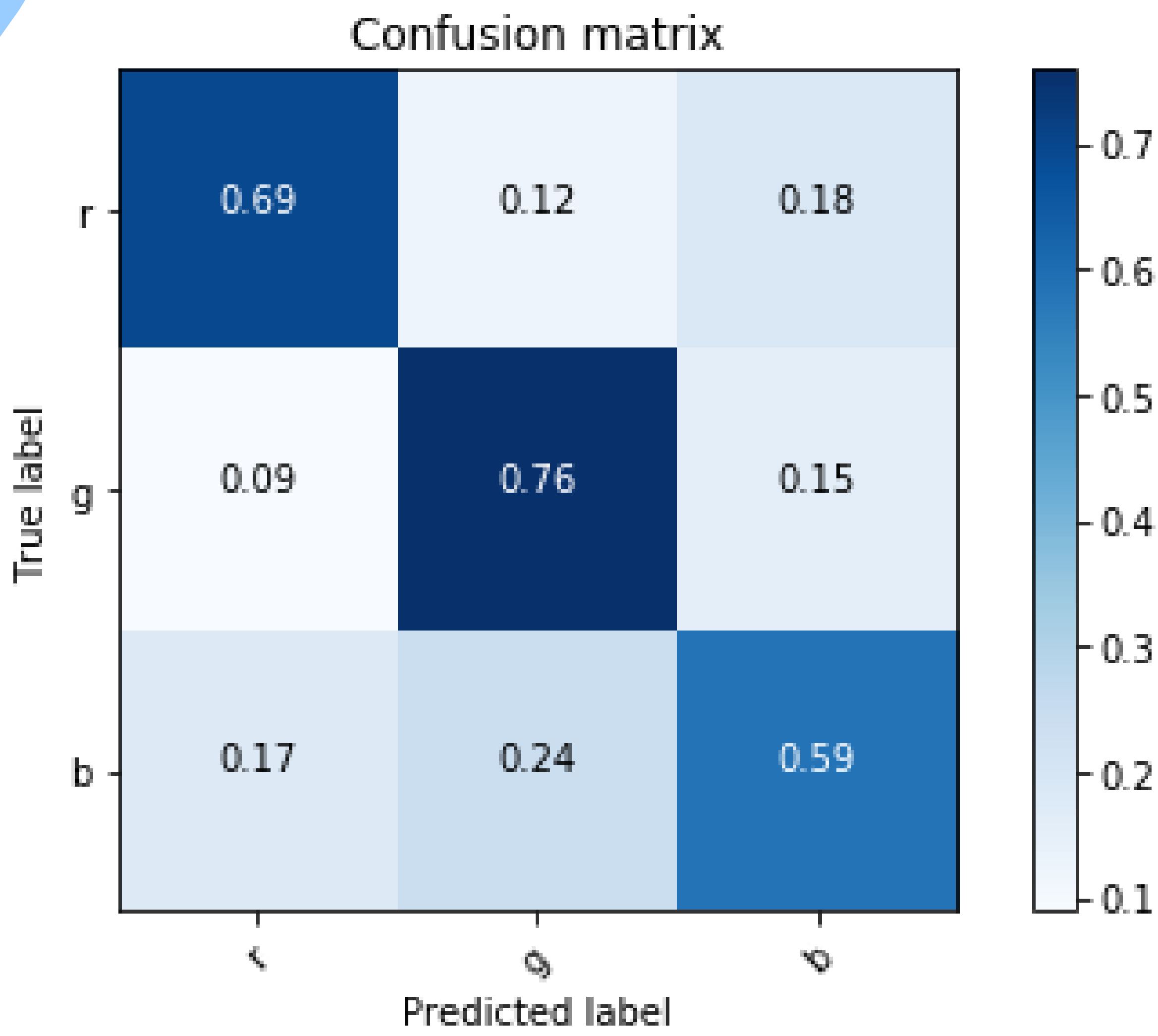
Labels may need to be flipped

- Confusion Matrix:
  - `cm = confusion_matrix(y,y_pred)`

Decision Boundary



Confusion Matrix



Gaussian Mixture Model

- Imports:
  - `from sklearn.mixture import GaussianMixture`
  - `from sklearn.metrics import confusion_matrix`
- Train:
  - `gmm = mixture.GaussianMixture(n_components=3)`
  - `gmm.fit(X)`

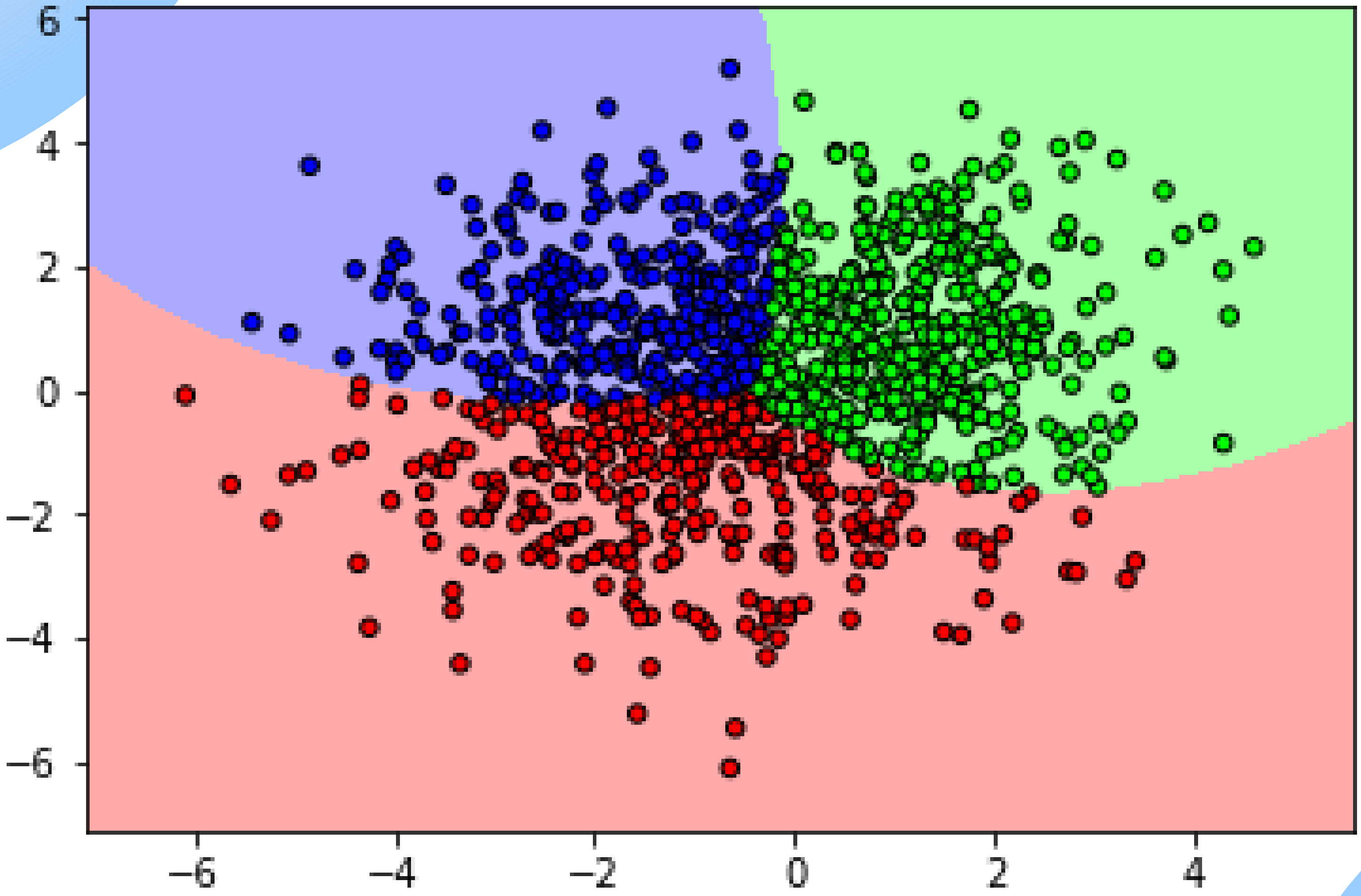
No-labels

- Predict:
  - `y_pred = gmm.predict(X)`

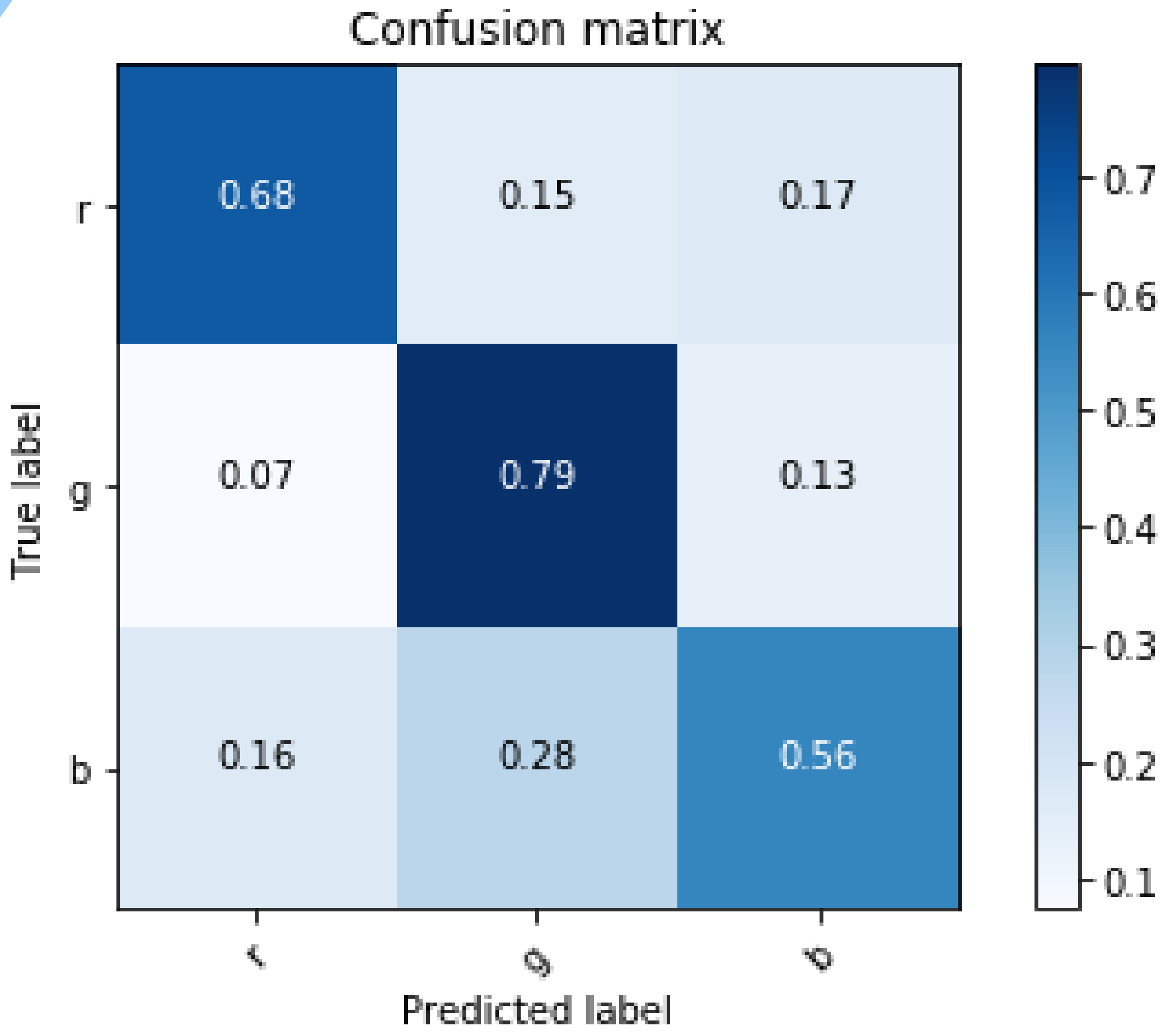
Labels may need to be flipped

- Confusion Matrix:
  - `cm = confusion_matrix(y,y_pred)`

Decision Boundary



Confusion Matrix



Conclussion

- Introduction to supervised classifiation (labelled data):
  - Generative: Naive-Bayes
  - Discriminative: Logistic Regression
- Introduction to unsupervised classification:
  - *k*-means
  - GMM