

Elements Of Data Science - F2024

Week 6: Intro to Machine Learning Models Continued

10/14/2024

TODOs

- Readings:
 - PDSH 05.03 Hyperparameters and Model Validation
 - Recommended: PML Chapter 6 (Except for Pipelines) and sklearn - model selection
 - Reference: PML Chapter Chap 3, 7, and sklearn - supervised learning

Today

- Review Linear Models
- Distance Based: kNN
- Tree Based: Decision Tree
- Ensembles: Bagging, Boosting, Stacking
- Multiclass/Multilabel and One Vs. Rest Classification
- Model Review

Questions?

Environment Setup

Environment Setup

```
In [1]: 1 import numpy as np
        2 import pandas as pd
        3 import matplotlib.pyplot as plt
        4 import seaborn as sns
        5
        6 from mlxtend.plotting import plot_decision_regions
        7
        8 from sklearn.linear_model import LinearRegression, LogisticRegression
        9
       10 sns.set_style('darkgrid')
       11 %matplotlib inline
```

Environment Setup

```
In [1]: 1 import numpy as np
        2 import pandas as pd
        3 import matplotlib.pyplot as plt
        4 import seaborn as sns
        5
        6 from mlxtend.plotting import plot_decision_regions
        7
        8 from sklearn.linear_model import LinearRegression, LogisticRegression
        9
       10 sns.set_style('darkgrid')
       11 %matplotlib inline
```

```
In [2]: 1 def my_plot_decision_regions(X,y,model,figsize=(5,5),ax=None):
        2     '''Plot classifier decision regions, classification predictions and training data'''
        3     if not ax:
        4         fig,ax = plt.subplots(1,1,figsize=figsize)
        5         # use mlxtend plot_decision_regions
        6         model = model.fit(X.values,y.values)
        7         plot_decision_regions(X.values,y.values,model,ax=ax)
        8         ax.set_xlabel(X.columns[0]); ax.set_ylabel(X.columns[1]);
        9
       10 def my_plot_regression(X,y,model,label='yhat',figsize=(5,5),ax=None):
       11     '''Plot regression predictions and training data'''
       12     # generate test data and make predictions
       13     X_test = np.linspace(X.iloc[:,0].min(),X.iloc[:,0].max(),1000).reshape(-1,1)
       14     model = model.fit(X.values,y.values)
       15     y_hat = model.predict(X_test)
       16     fig,ax = plt.subplots(1,1,figsize=figsize)
       17     ax.scatter(X, y, s=20, edgecolor="black", c="darkorange", label="data")
       18     ax.plot(X_test, y_hat, color="cornflowerblue", label=label, linewidth=2)
       19     ax.set_xlabel(X.columns[0]); ax.set_ylabel(y.name); ax.legend();
```

Linear Models (Review)

- Simple/Multiple Linear Regression
- Logistic Regression
- SVM
- Perceptron, Multi-Layer Perceptron

Wine as Binary Classification

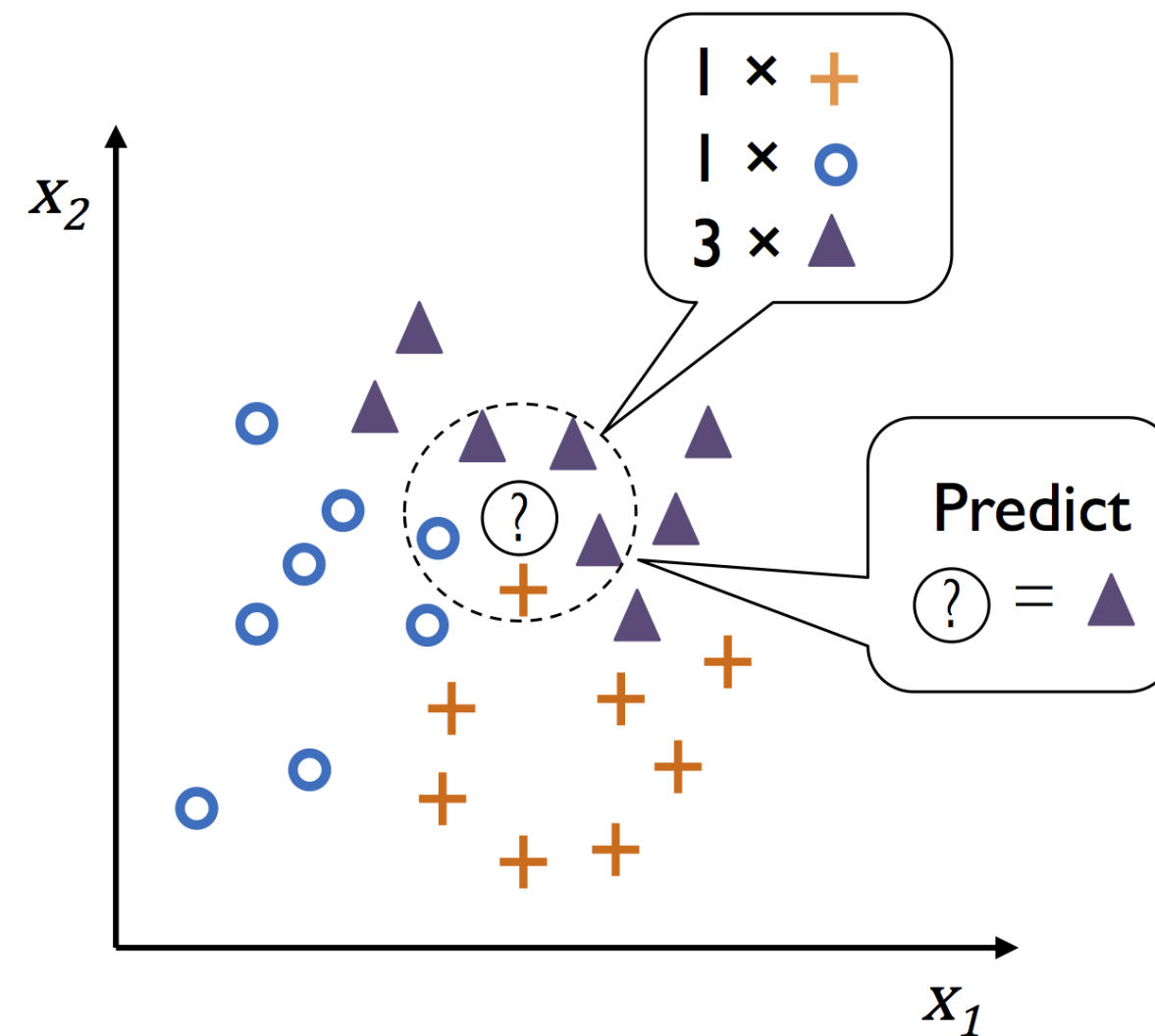
Wine as Binary Classification

```
In [3]: 1 df_wine = pd.read_csv('../data/wine_dataset.csv',usecols=['alcohol','ash','proline','hue','class'])
2 # rename 'class' as 'target', since class is a reserved python word
3 df_wine = df_wine.rename({'class':'target'},axis=1)
4
5 df_wine_2c = df_wine[df_wine.target < 2] # only keep classes 0 and 1
6
7 X_2c = df_wine_2c[['proline','hue']]
8 y_2c = df_wine_2c['target']
9
10 zscore = lambda x: (x-x.mean()) / x.std()
11
12 X_2c_zscore = X_2c.apply(zscore,axis=0)
13 alcohol_2c_zscore = zscore(df_wine_2c.alcohol)
14
15 y_2c.value_counts().sort_index()
```

```
Out[3]: 0    59
1    71
Name: target, dtype: int64
```

Distance Based: k-Nearest Neighbor (kNN)

- What category do most of the k nearest neighbors belong to?

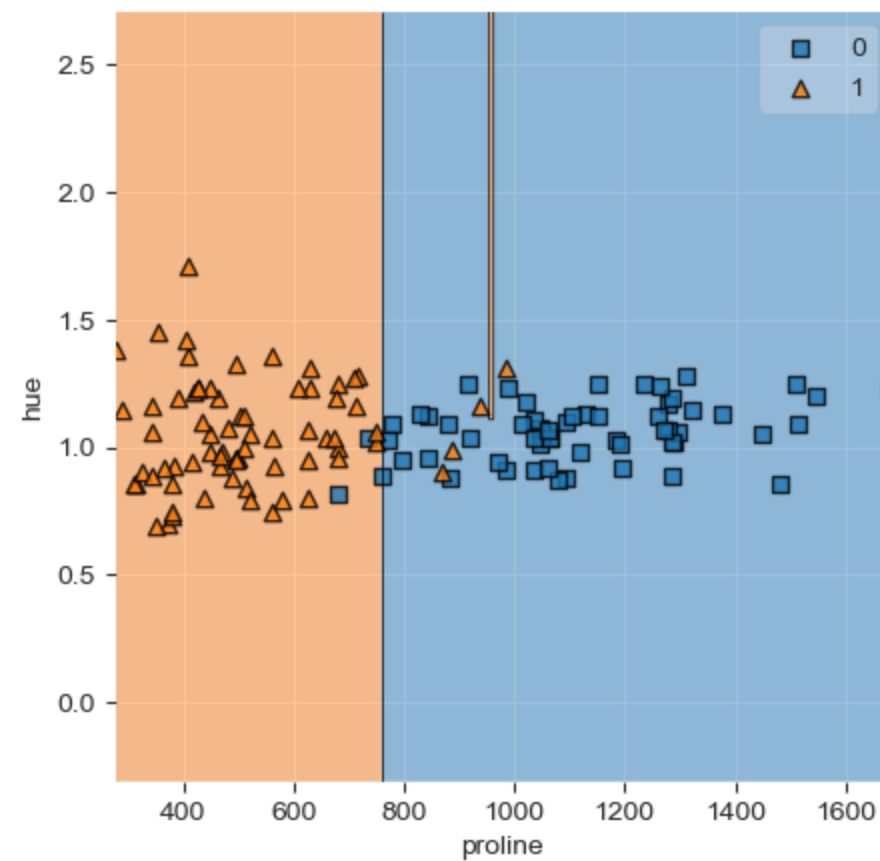


From PML

KNN in sklearn

KNN in sklearn

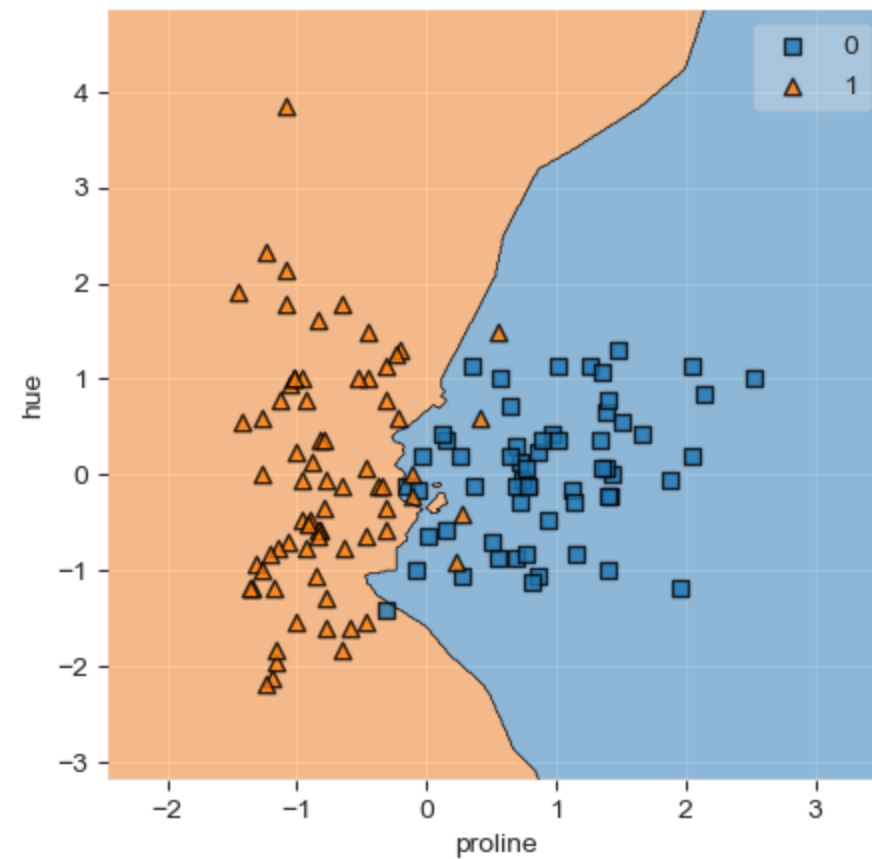
```
In [4]: 1 from sklearn.neighbors import KNeighborsClassifier
2
3 knn = KNeighborsClassifier(n_neighbors=3)
4 knn.fit(X_2c,y_2c)
5
6 my_plot_decision_regions(X_2c,y_2c,knn)
```



Effects of Standardization on Distance Based Methods

Effects of Standardization on Distance Based Methods

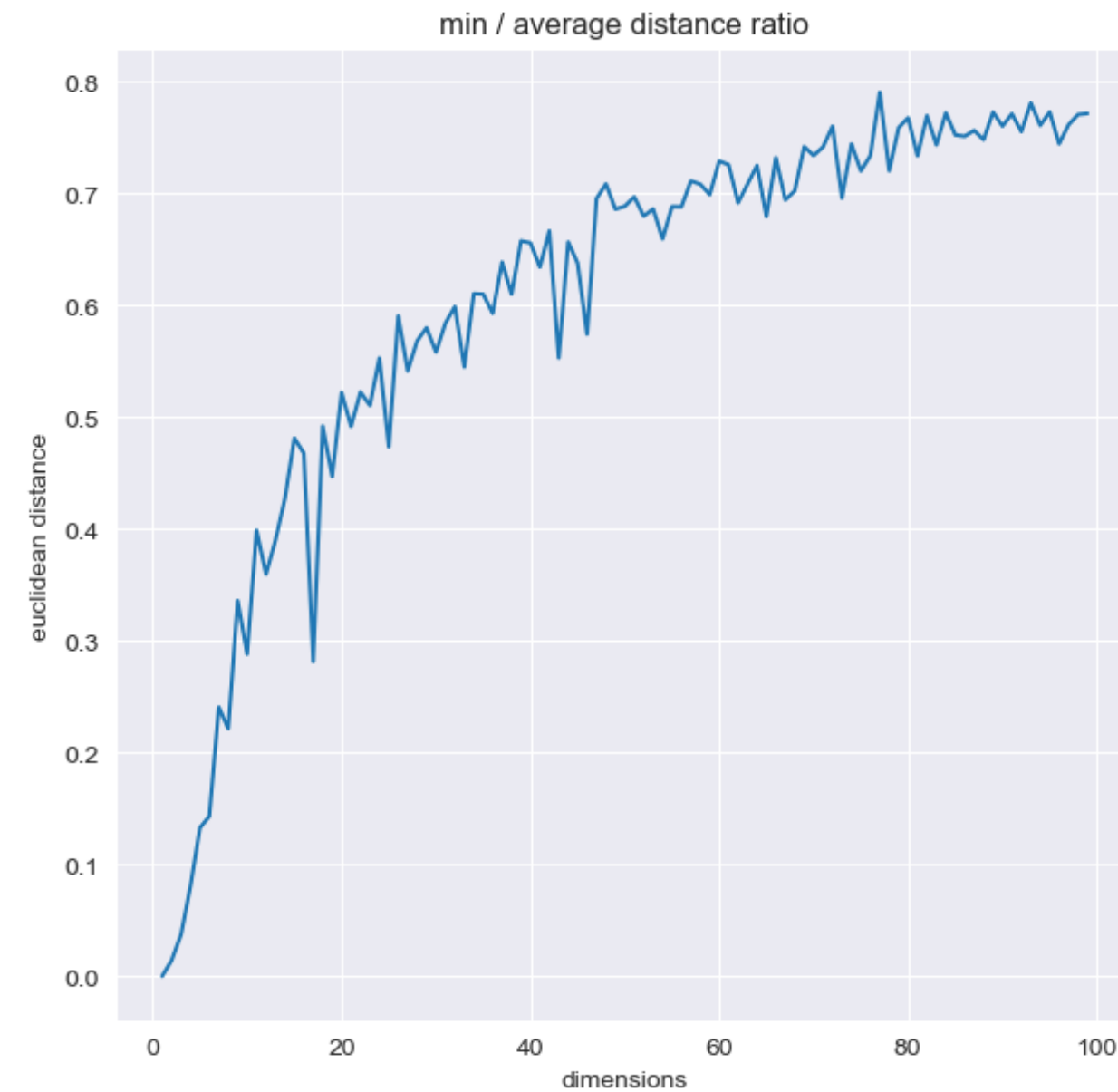
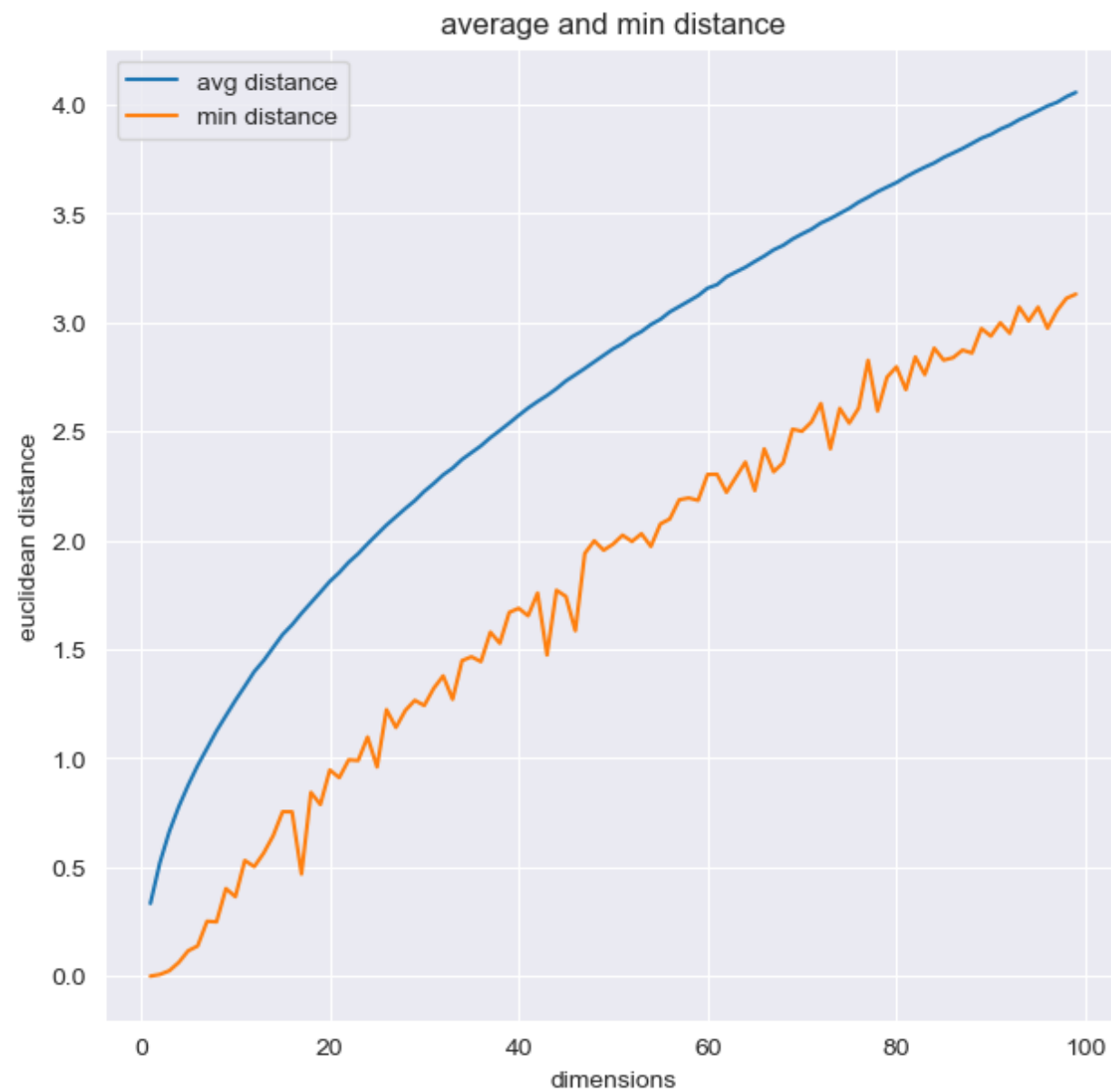
```
In [5]: 1 knn_z = KNeighborsClassifier(n_neighbors=3)
2 knn_z.fit(X_2c_zscore,y_2c)
3
4 my_plot_decision_regions(X_2c_zscore,y_2c,knn_z)
```



Curse of Dimensionality Cont.

Curse of Dimensionality Cont.

```
In [7]: 1 fig,ax = plt.subplots(1,2,figsize=(16,7))
2 ax[0].plot(dimensions,avg_distances,label='avg distance');
3 ax[0].plot(dimensions,min_distances,label='min distance');
4 ax[0].legend()
5 ax[0].set_title('average and min distance'); ax[0].set_xlabel('dimensions'); ax[0].set_ylabel('euclidean distance');
6 ax[1].plot(dimensions,min_avg_ratio)
7 ax[1].set_title('min / average distance ratio'); ax[1].set_xlabel('dimensions'); ax[1].set_ylabel('euclidean distance');
```



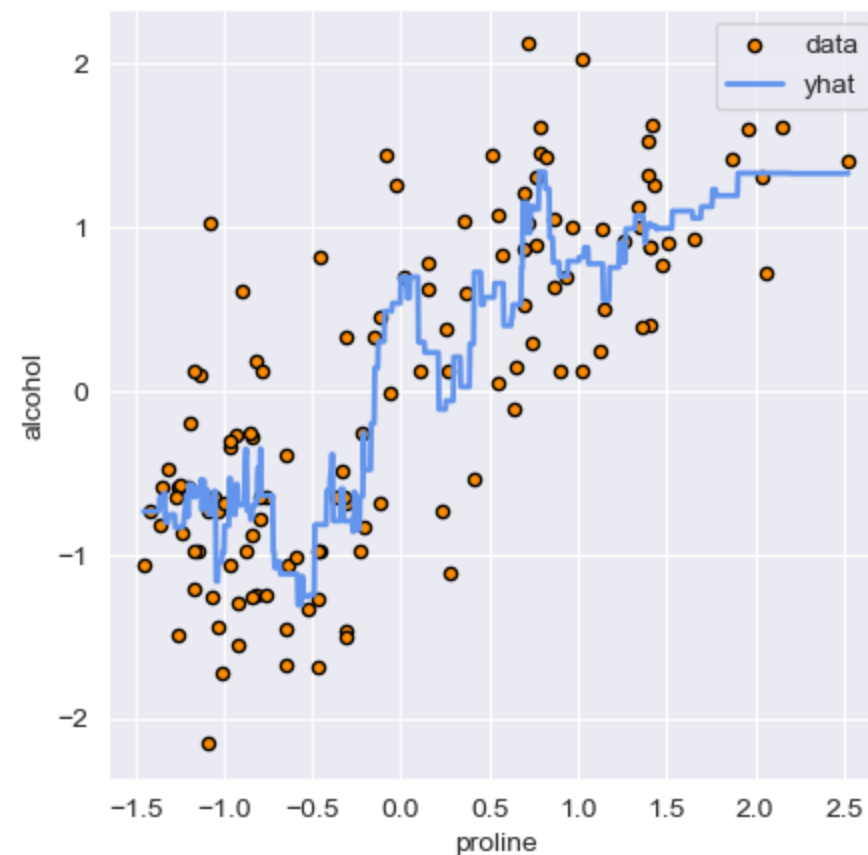
Regression with kNN

Approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood.

Regression with kNN

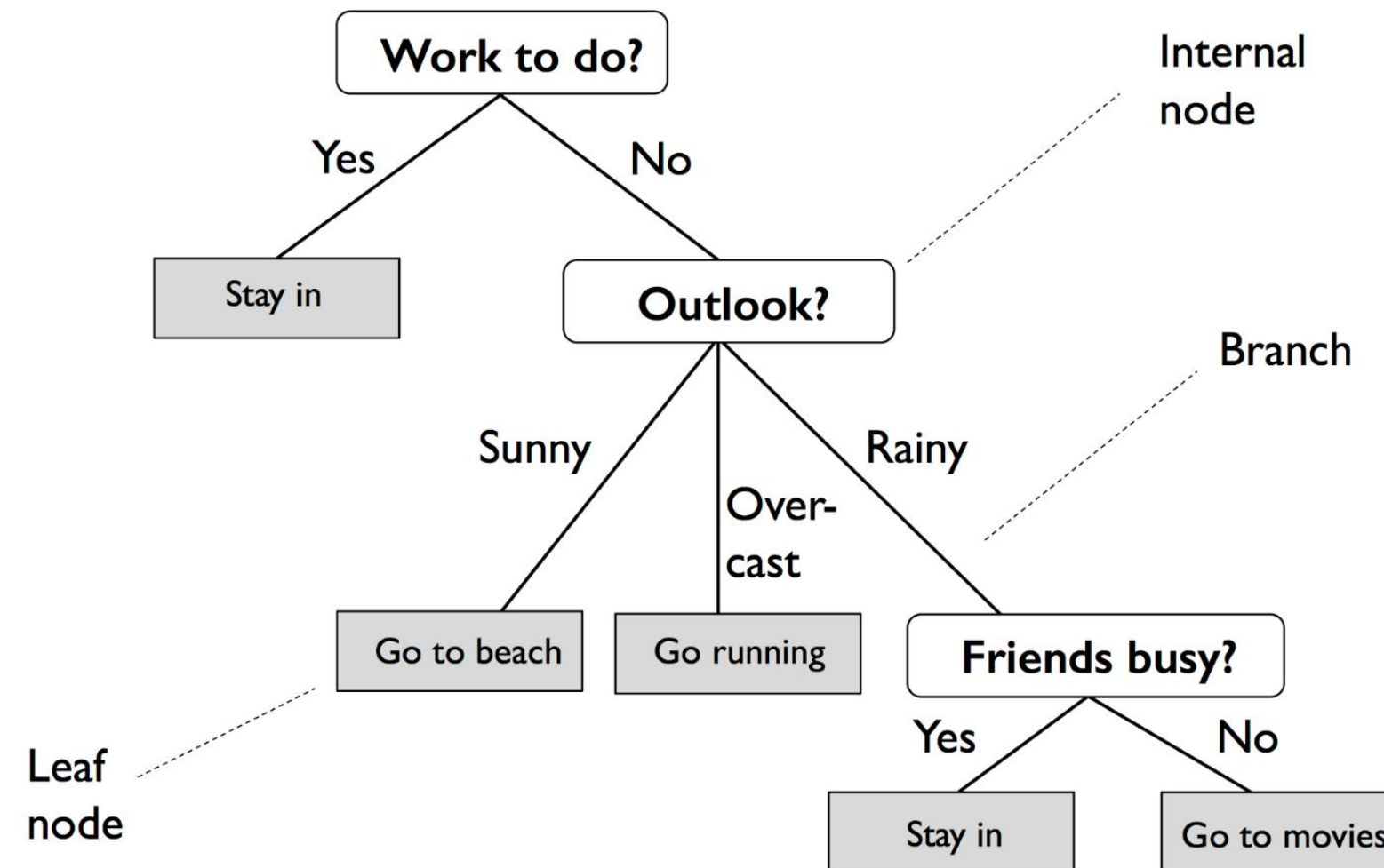
Approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood.

```
In [8]: 1 from sklearn.neighbors import KNeighborsRegressor
2
3 knnr = KNeighborsRegressor(n_neighbors=5)
4 knnr.fit(X_2c_zscore[['proline']], alcohol_2c_zscore)
5
6 my_plot_regression(X_2c_zscore[['proline']], alcohol_2c_zscore, knnr)
```



Decision Tree

- What answer does a series of yes/no questions lead us to?

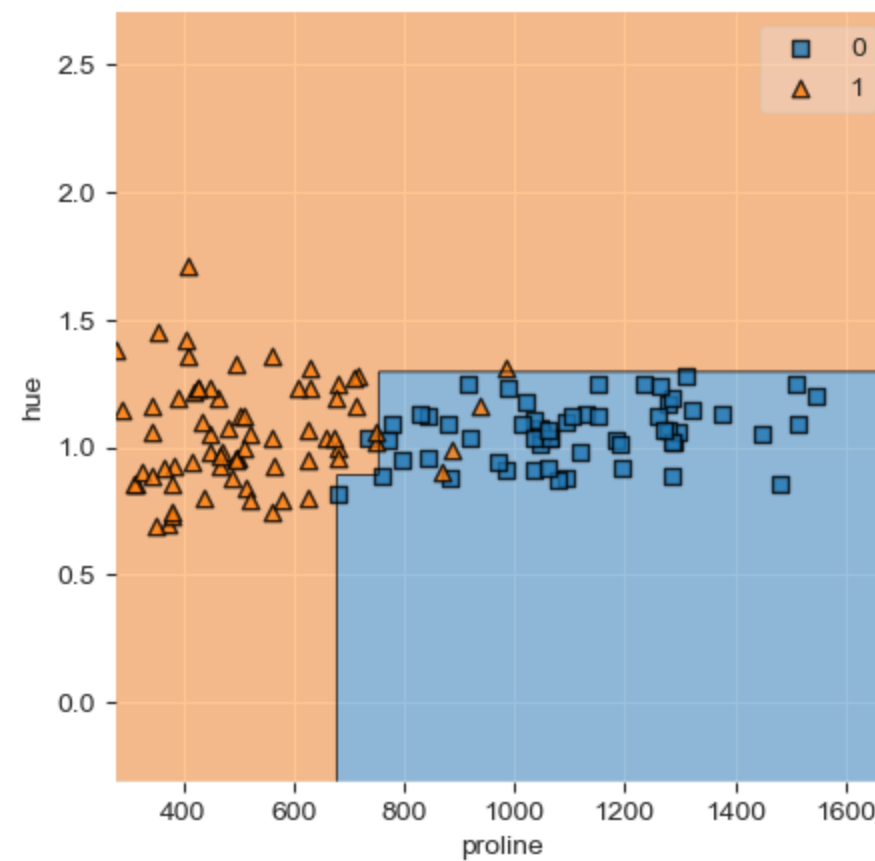


From PML

Decision Tree Classifier in sklearn

Decision Tree Classifier in sklearn

```
In [10]: 1 from sklearn.tree import DecisionTreeClassifier
2
3 dtc_md3 = DecisionTreeClassifier(max_depth=3) # max_depth: max number of questions
4 dtc_md3.fit(X_2c,y_2c)
5
6 my_plot_decision_regions(X_2c,y_2c,dtc_md3)
```

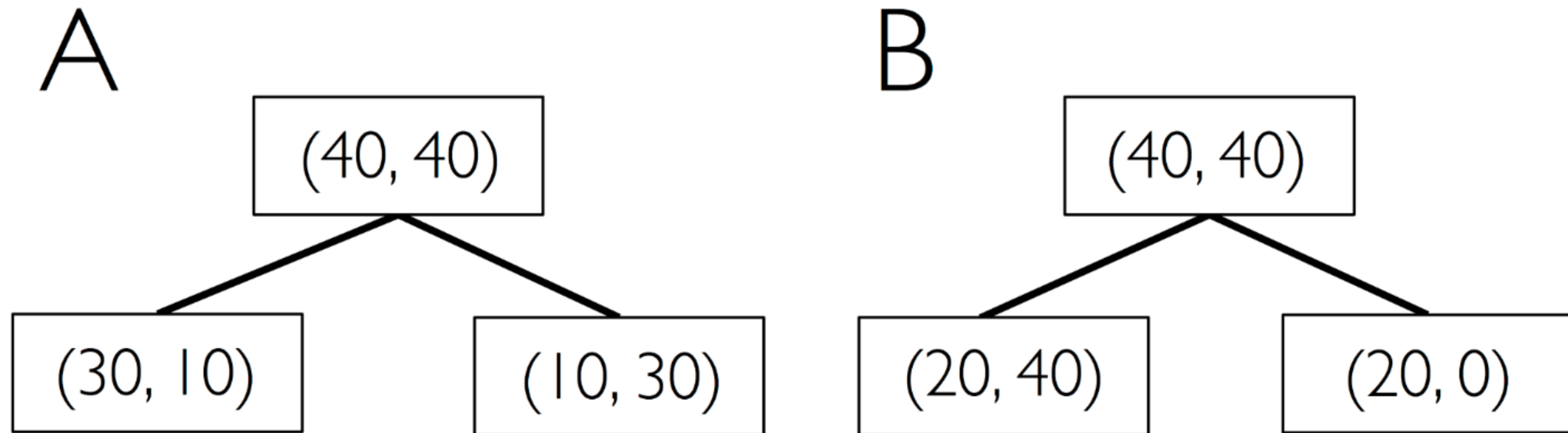


Building a Decision Tree

- How to decide which question to choose (eg. Should I choose question A or B)?
- **Reduce Impurity**

Building a Decision Tree

- How to decide which question to choose (eg. Should I choose question A or B)?
- **Reduce Impurity**



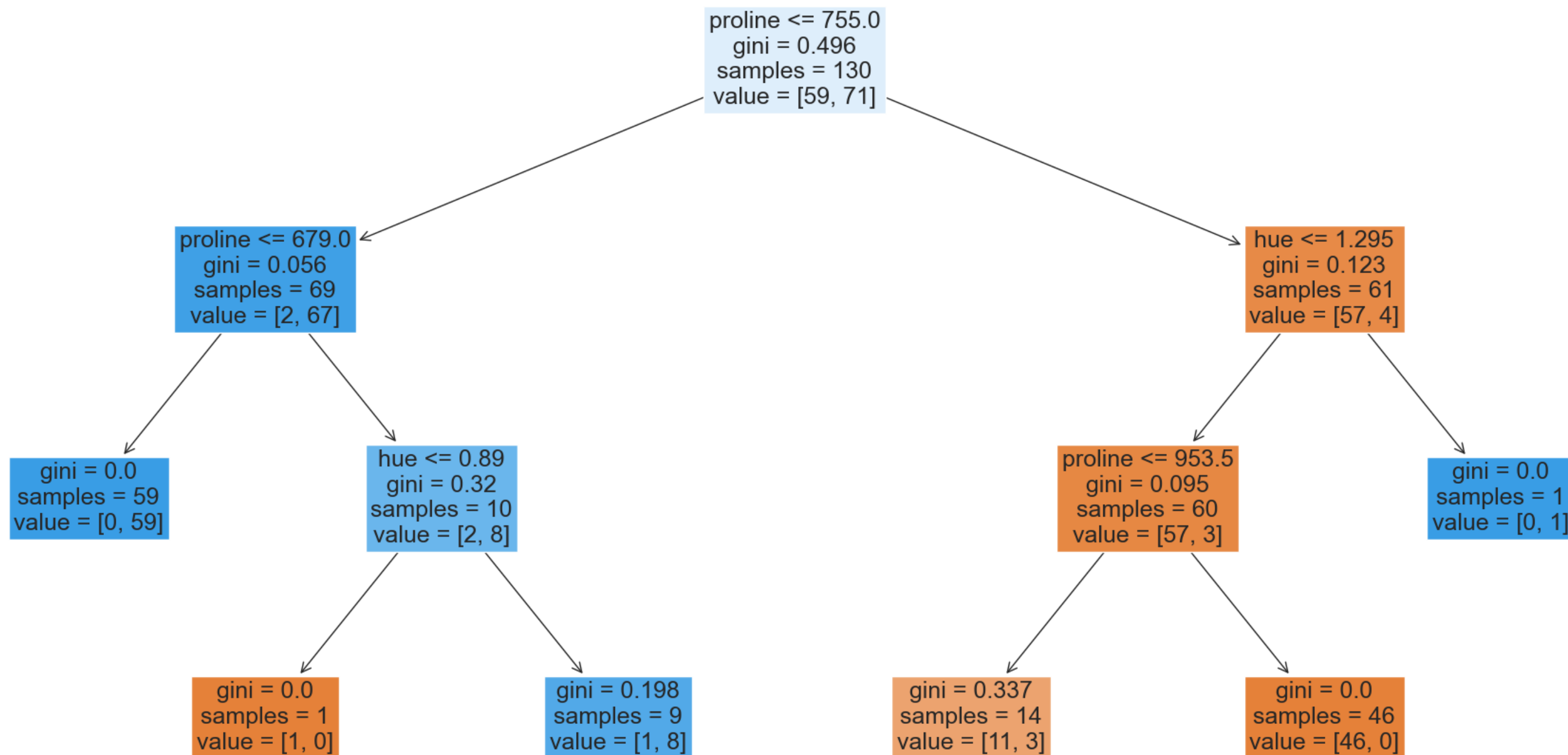
From PML

- Information Gain: Tie, *Gini*: B, Entropy: B

Plot Learned Decision Tree Using sklearn

Plot Learned Decision Tree Using sklearn

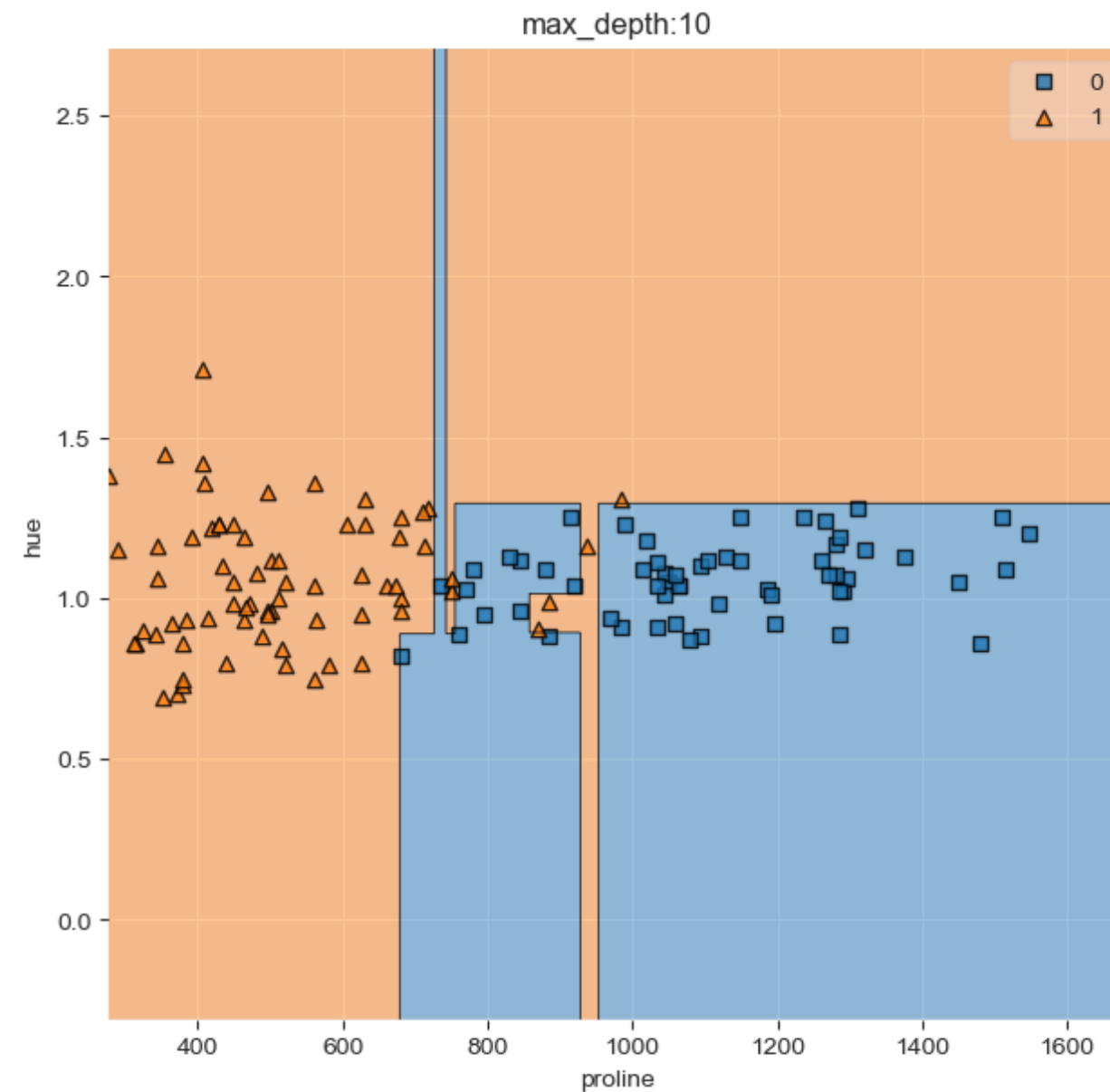
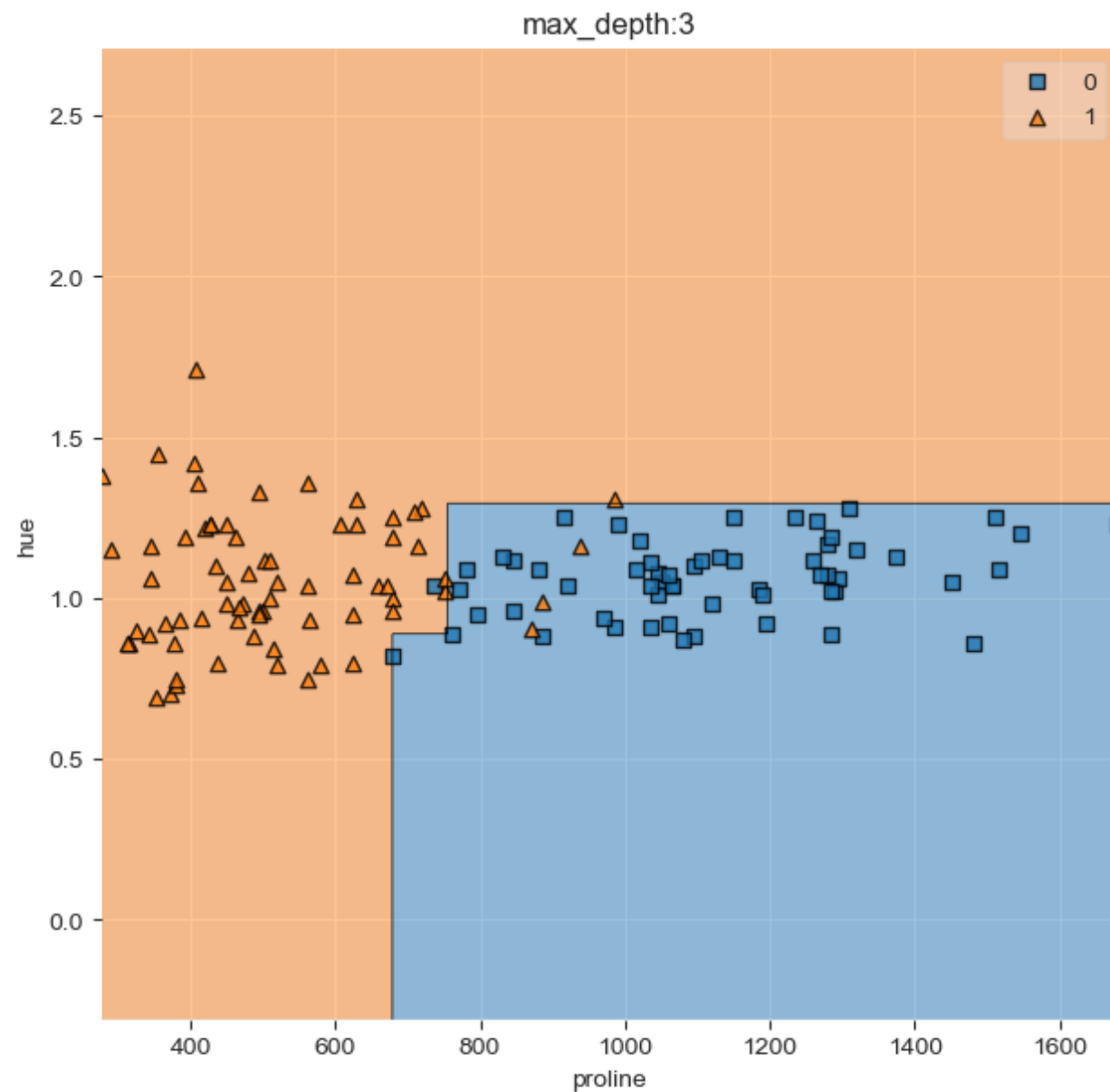
```
In [12]: 1 from sklearn.tree import plot_tree
2 fig, ax = plt.subplots(1, 1, figsize=(24, 12))
3 plot_tree(dtc_md3, ax=ax, fontsize=18, feature_names=X_2c.columns, filled=True);
```



Decision Tree: Increase Maximum Depth

Decision Tree: Increase Maximum Depth

```
In [13]: 1 dtc_md10 = DecisionTreeClassifier(max_depth=10)
2 dtc_md10.fit(X_2c,y_2c)
3
4 fig,ax = plt.subplots(1,2,figsize=(16,7))
5 my_plot_decision_regions(X_2c, y_2c, model=dtc_md3, ax=ax[0]);
6 my_plot_decision_regions(X_2c, y_2c, model=dtc_md10, ax=ax[1]);
7 ax[0].set_title('max_depth:3');ax[1].set_title('max_depth:10');
```



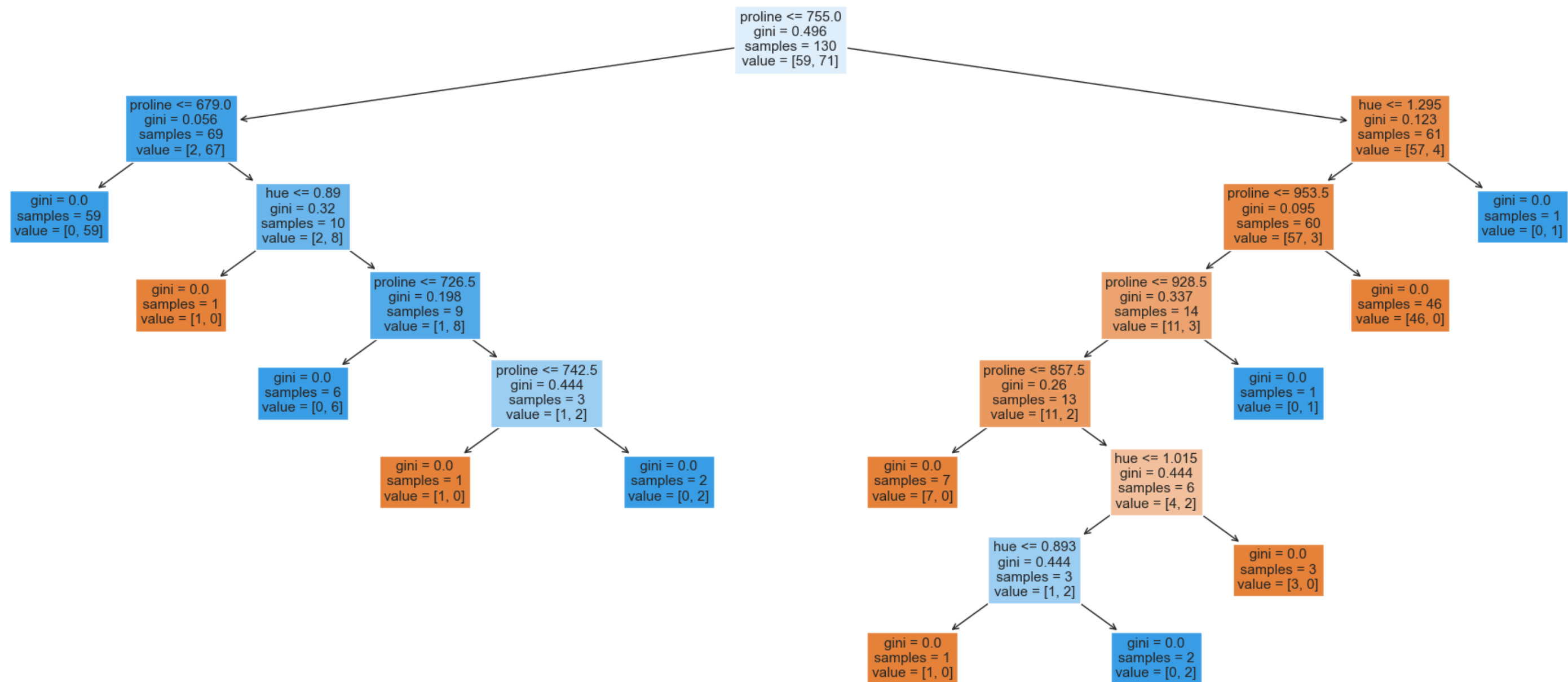
Plot Learned Decision Tree Using sklearn

- For tree with max_depth=10

Plot Learned Decision Tree Using sklearn

- For tree with max_depth=10

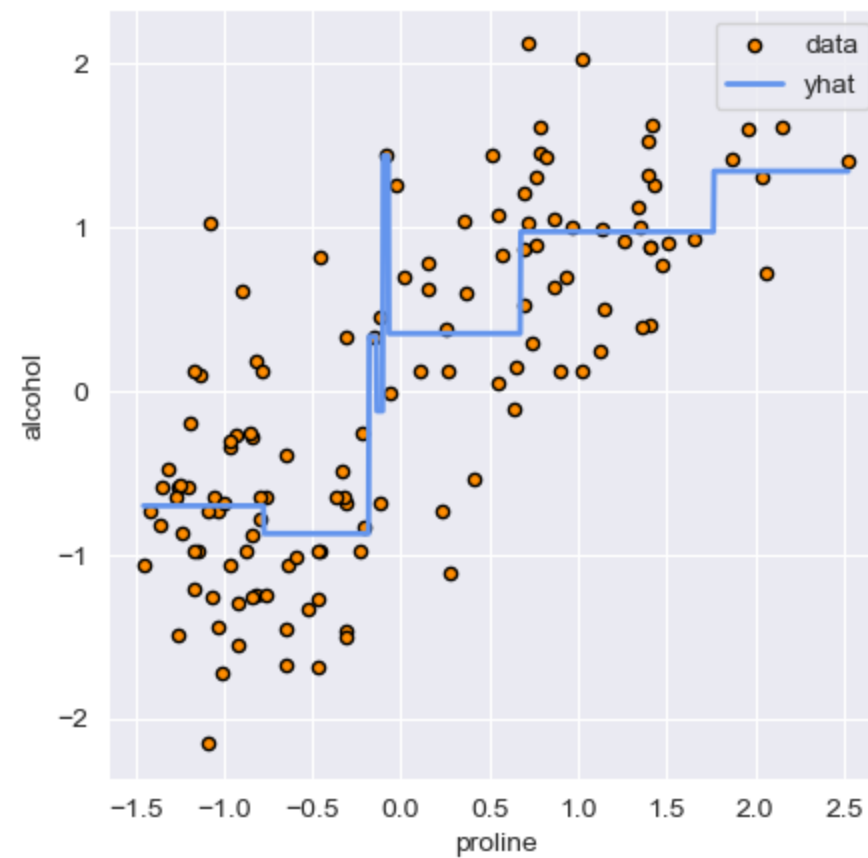
```
In [14]: 1 fig,ax = plt.subplots(1,1,figsize=(24,10))
2 plot_tree(dtc_md10,ax=ax,fontsize=11,feature_names=X_2c.columns,filled=True);
```



Regression with Decision Trees

Regression with Decision Trees

```
In [15]: 1 from sklearn.tree import DecisionTreeRegressor
2
3 dtr = DecisionTreeRegressor(max_depth=3)
4 dtr.fit(X_2c_zscore[['proline']], alcohol_2c_zscore)
5
6 my_plot_regression(X_2c_zscore[['proline']], alcohol_2c_zscore, dtr)
```



Ensemble Methods

- "Wisdom of the crowd"
- Can often achieve better performance with collection of learners
- Often use shallow trees as base learners

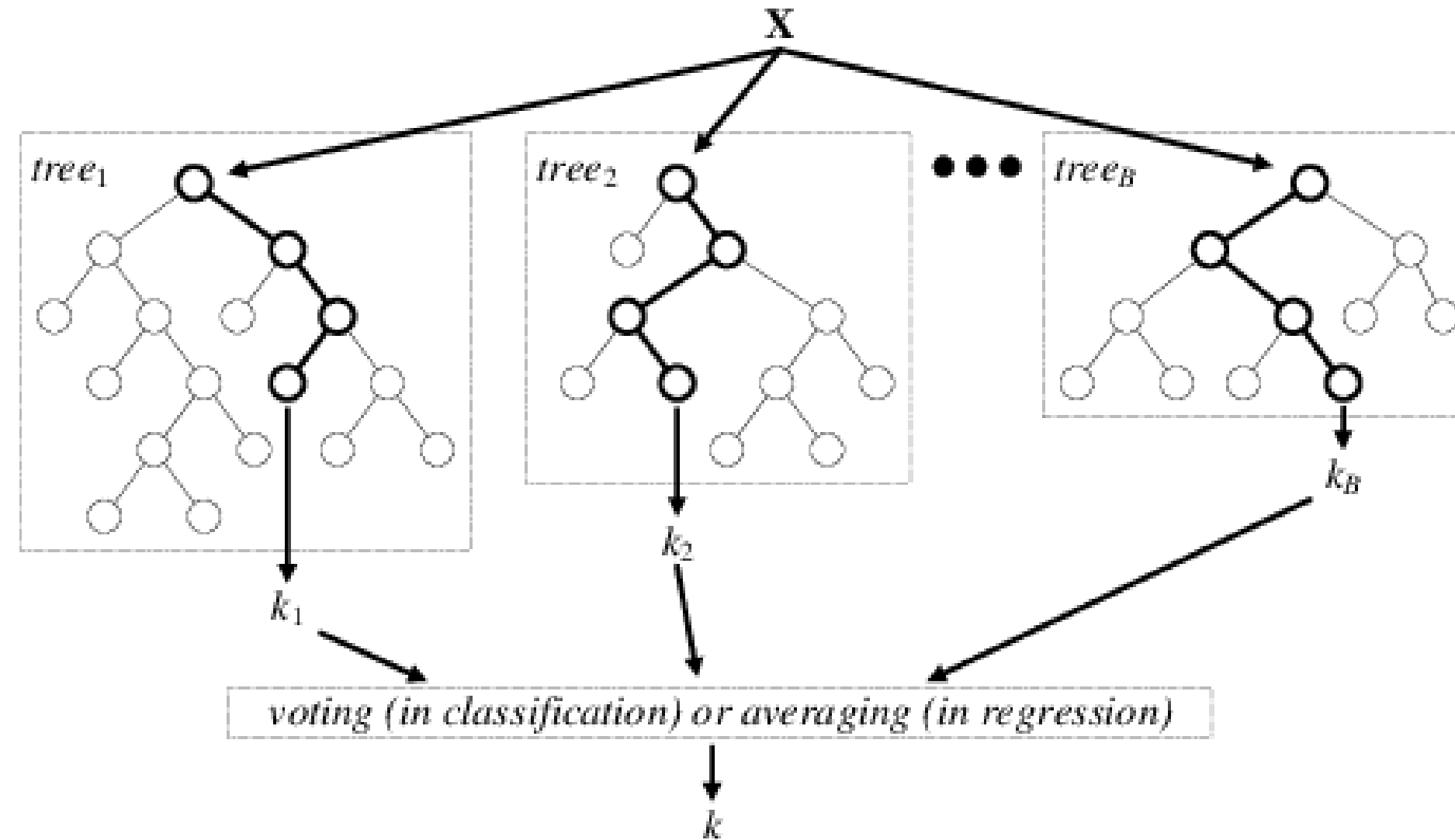
Ensemble Methods

- "Wisdom of the crowd"
- Can often achieve better performance with collection of learners
- Often use shallow trees as base learners

Common methods for generating ensembles:

- **Bagging** (Bootstrap Aggregation)
 - Random Forest
- **Boosting**
 - Gradient Boosting
- **Stacking**

Random Forest and Gradient Boosted Trees

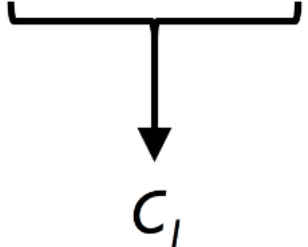


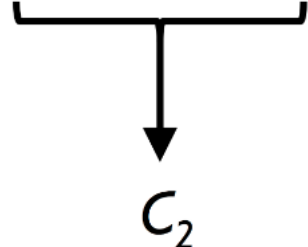
From https://www.researchgate.net/publication/301638643_Electromyographic_Patterns_during_Golf_Swing_Activation_Sequence_Profiling_and_Prediction_of_Shot_Effectiveness

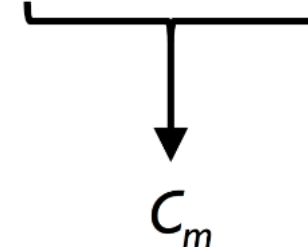
Bagging with Random Forests

- Trees built with bootstrap samples and subsets of features
- Achieve variation with random selection of observations and features

Sample indices	Bagging round 1	Bagging round 2	...
1	2	7	...
2	2	3	...
3	1	2	...
4	3	1	...
5	7	1	...
6	2	7	...
7	4	7	...

 C_1

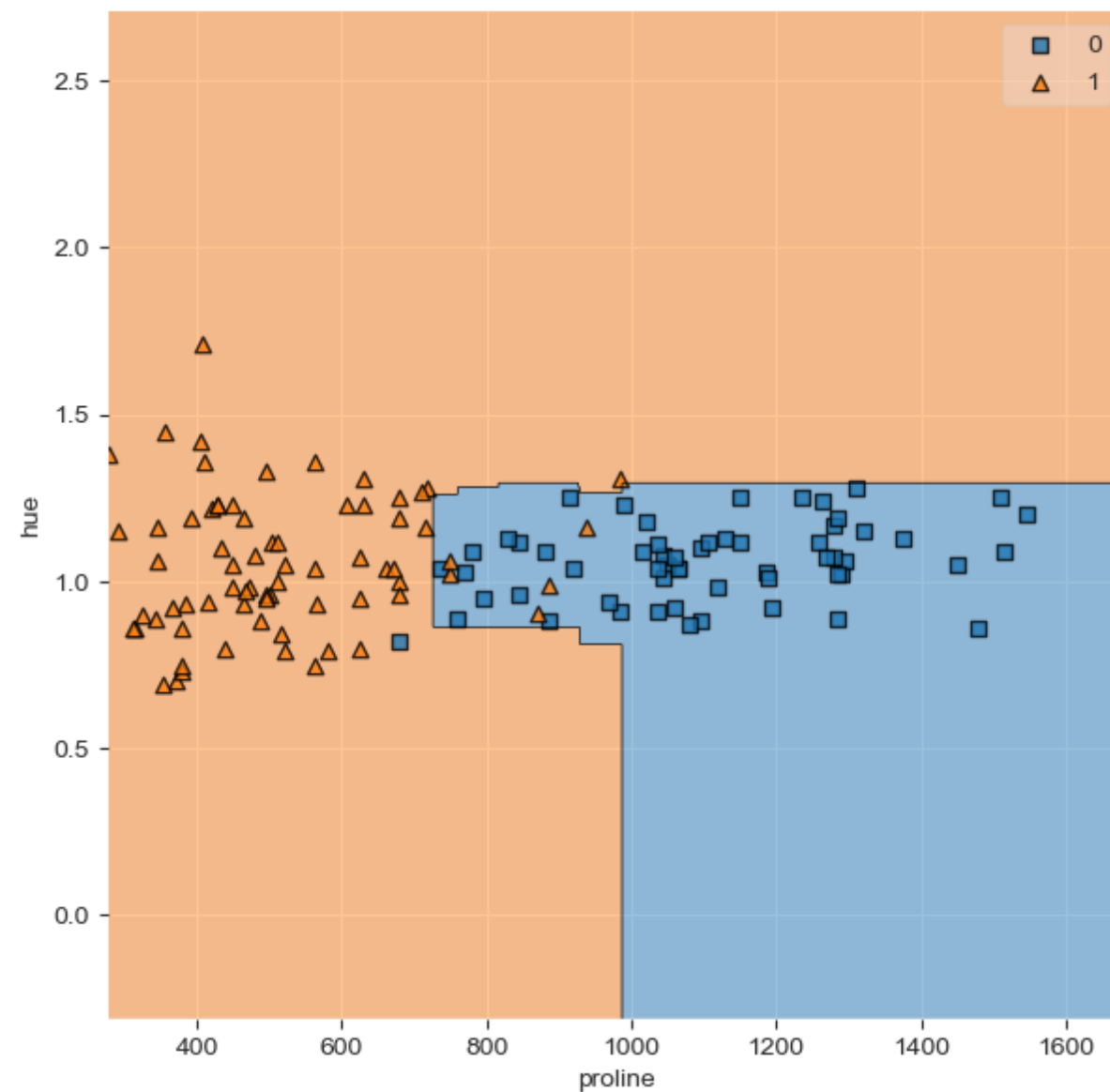
 C_2

 C_m

Random Forests with sklearn

Random Forests with sklearn

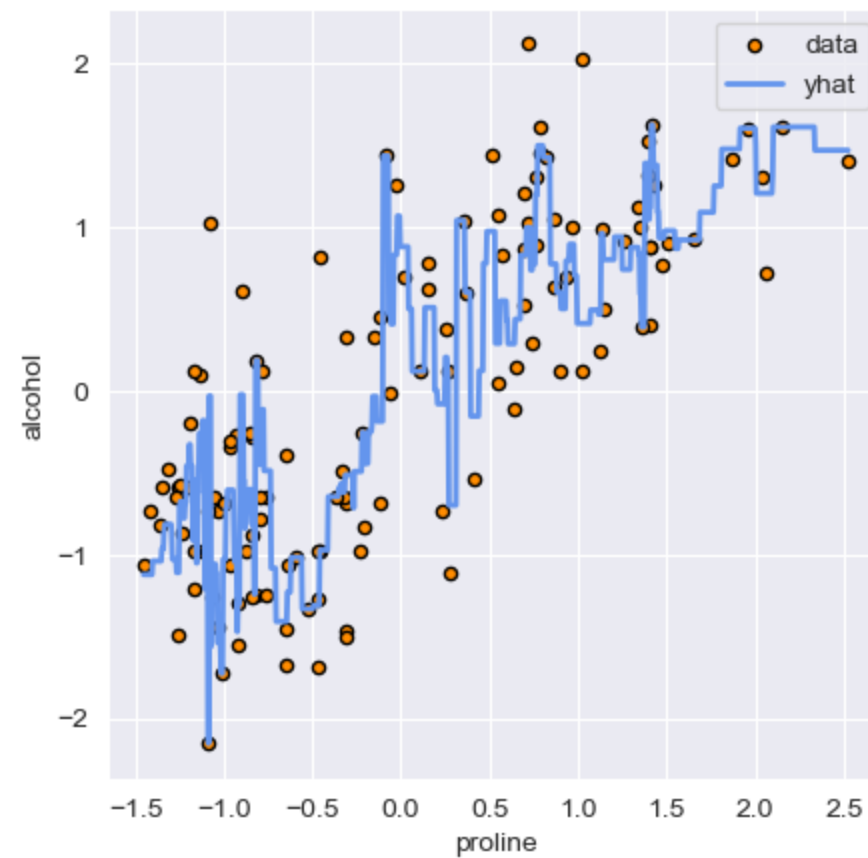
```
In [16]: 1 from sklearn.ensemble import RandomForestClassifier
2
3 rfc = RandomForestClassifier(n_estimators=10, # number of trees in ensemble
4                             max_depth=3,    # same as decision trees
5                             n_jobs=-1,      # parallelize using all available cores, default: None=1
6                             random_state=0) # for demonstration only
7 rfc.fit(X_2c,y_2c)
8 my_plot_decision_regions(X_2c,y_2c,rfc,figsize=(7,7))
```



Regression with RandomForest

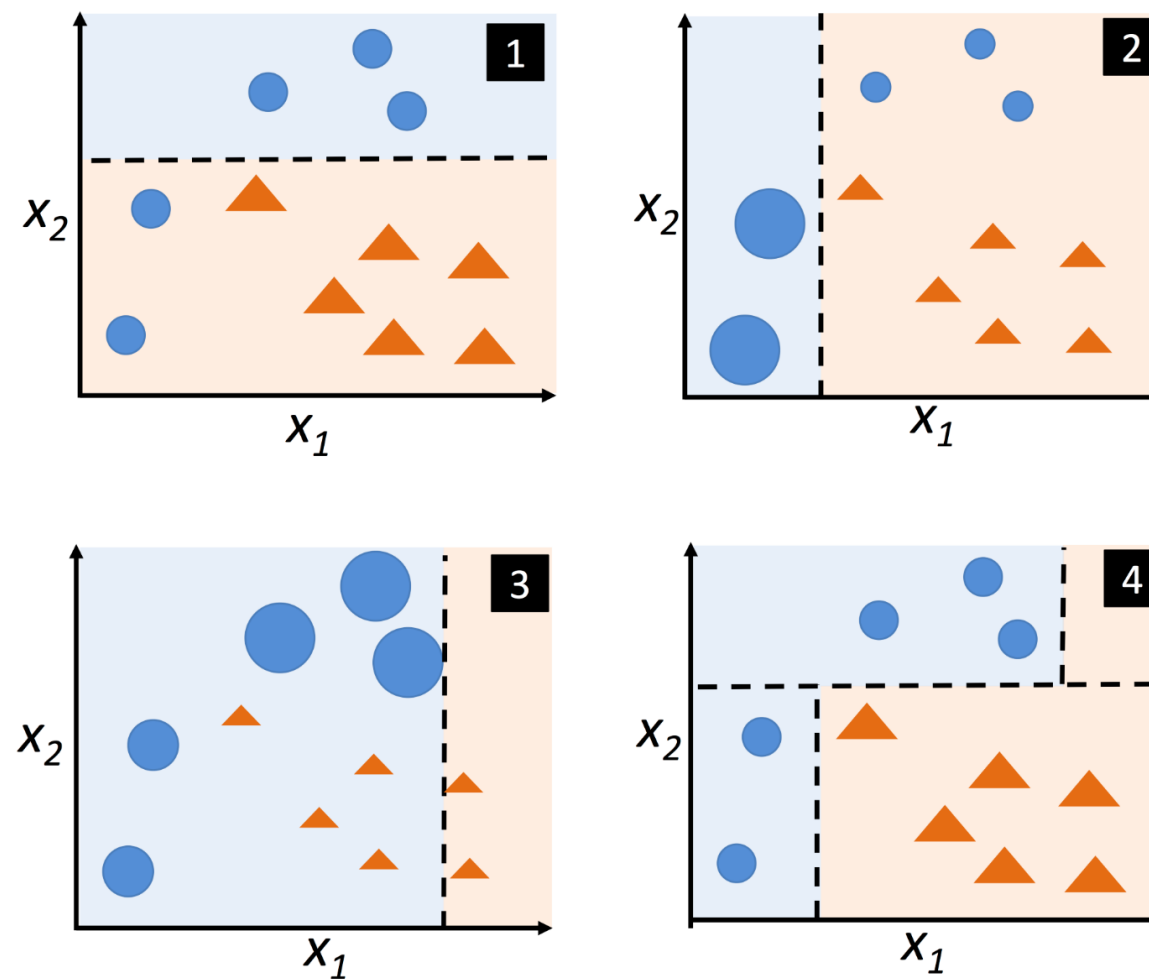
Regression with RandomForest

```
In [17]: 1 from sklearn.ensemble import RandomForestRegressor
2
3 rfr = RandomForestRegressor(n_estimators=3, n_jobs=-1)
4 rfr.fit(df_wine[['proline']],df_wine.alcohol)
5
6 my_plot_regression(X_2c_zscore[['proline']],alcohol_2c_zscore,rfr)
```



Gradient Boosted Trees

- Trees built by adding weight to mis-classification
- Achieve variation due to changes in weights on observations

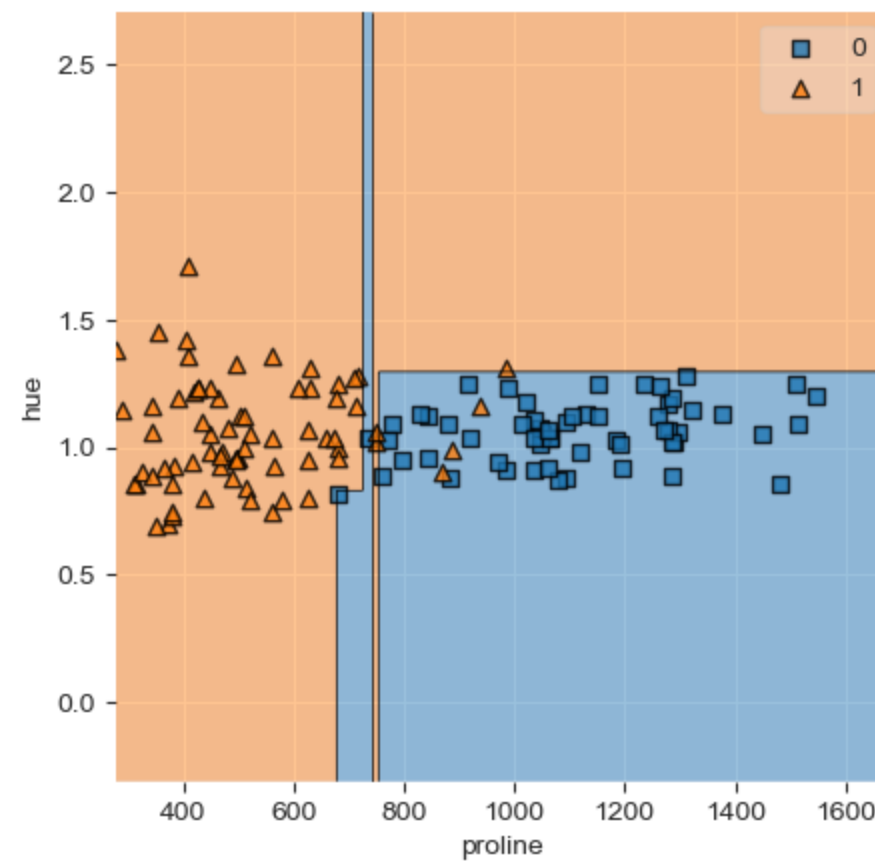


From PML

Gradient Boosted Trees in sklearn

Gradient Boosted Trees in sklearn

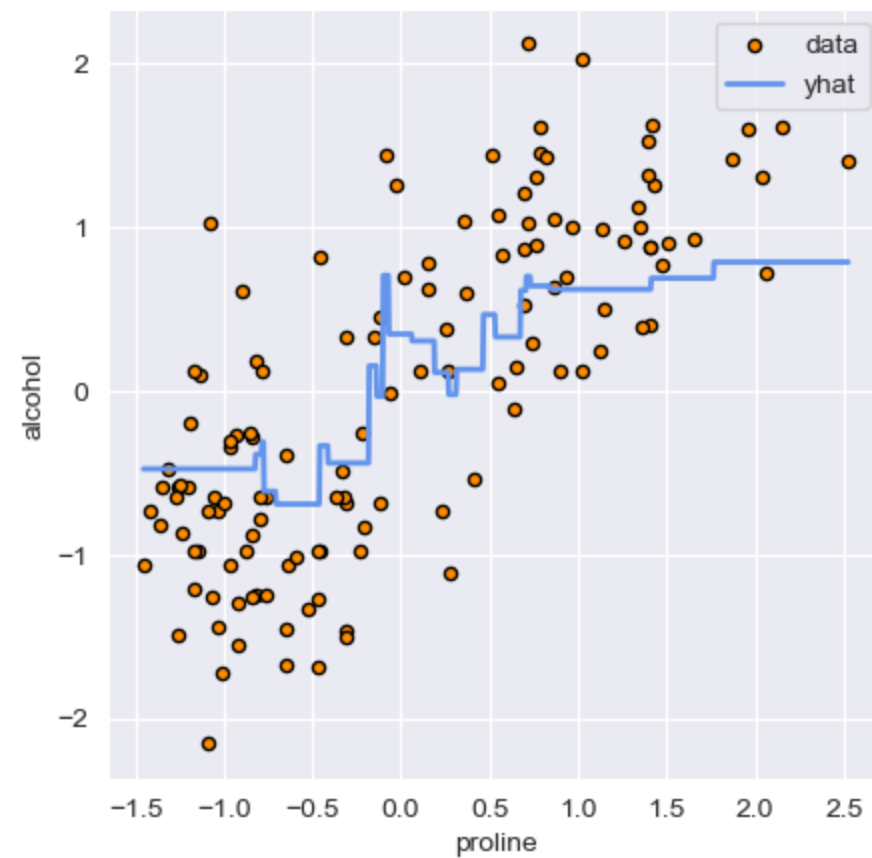
```
In [18]: 1 from sklearn.ensemble import GradientBoostingClassifier
2
3 gbc = GradientBoostingClassifier(n_estimators=10)
4 gbc.fit(X_2c,y_2c)
5
6 my_plot_decision_regions(X_2c,y_2c,gbc)
```



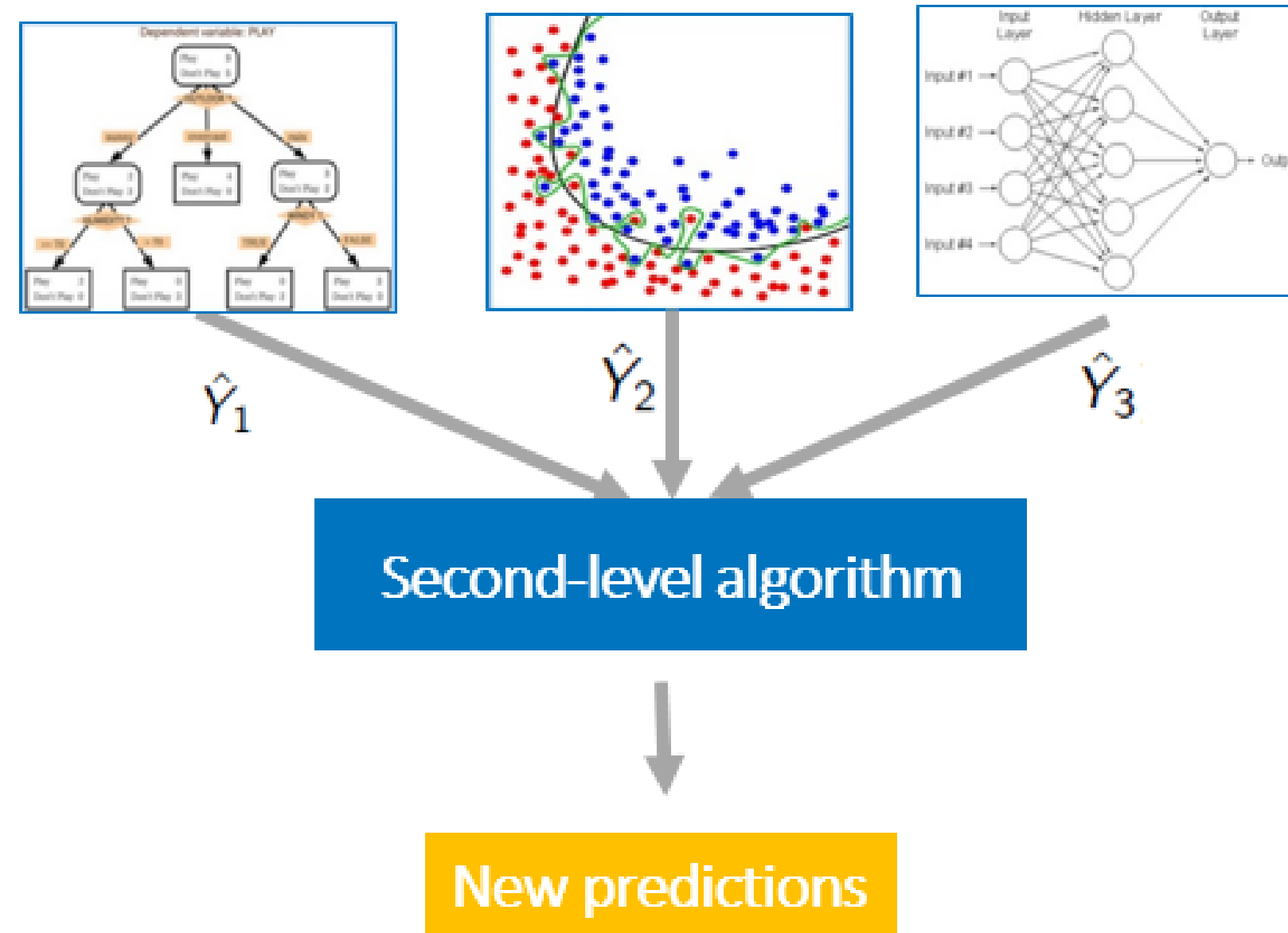
Regression with Gradient Boosted Trees

Regression with Gradient Boosted Trees

```
In [19]: 1 from sklearn.ensemble import GradientBoostingRegressor
2
3 gbr = GradientBoostingRegressor(n_estimators=10)
4 gbr.fit(X_2c_zscore[['proline']], alcohol_2c_zscore)
5
6 my_plot_regression(X_2c_zscore[['proline']], alcohol_2c_zscore, gbr)
```



Stacking

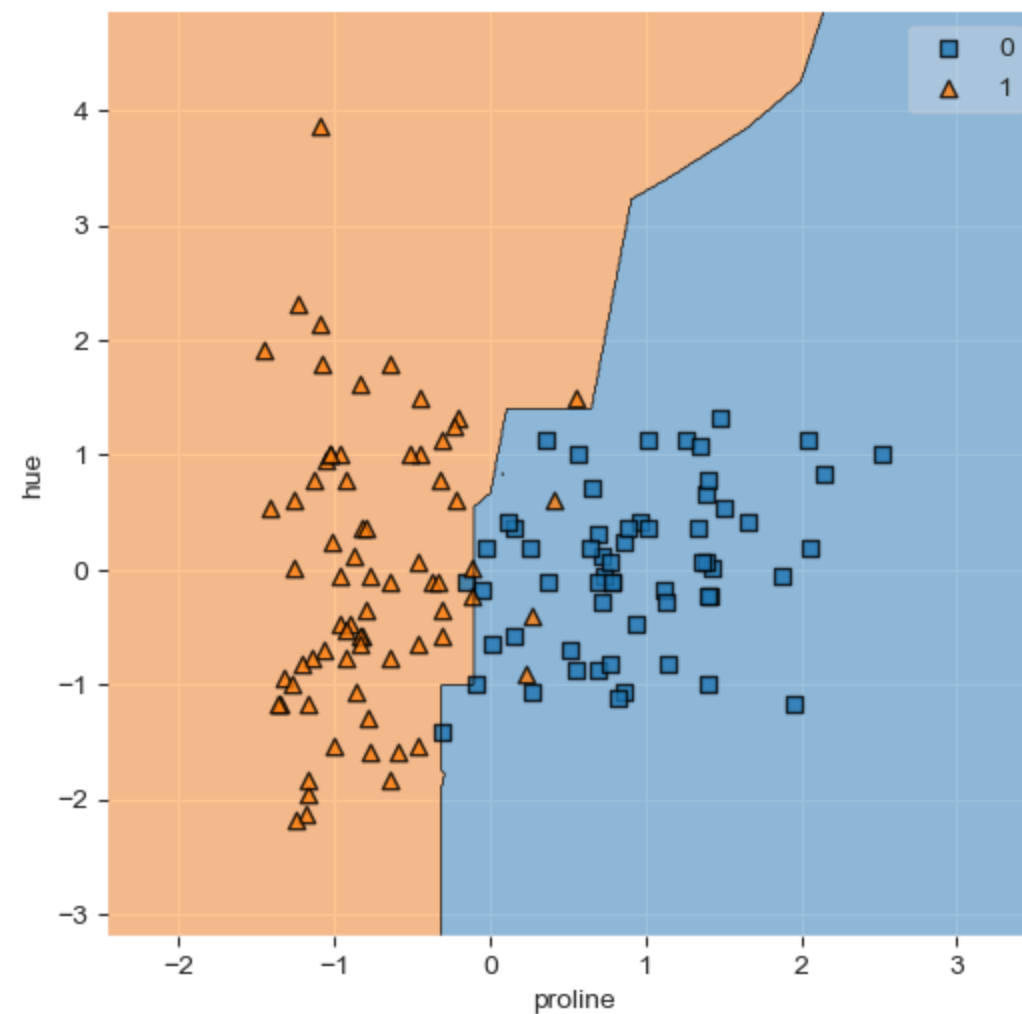


From <https://blogs.sas.com/content/subconsciousmusings/2017/05/18/stacked-ensemble-models-win-data-science-competitions/>

Stacking for Classification

Stacking for Classification

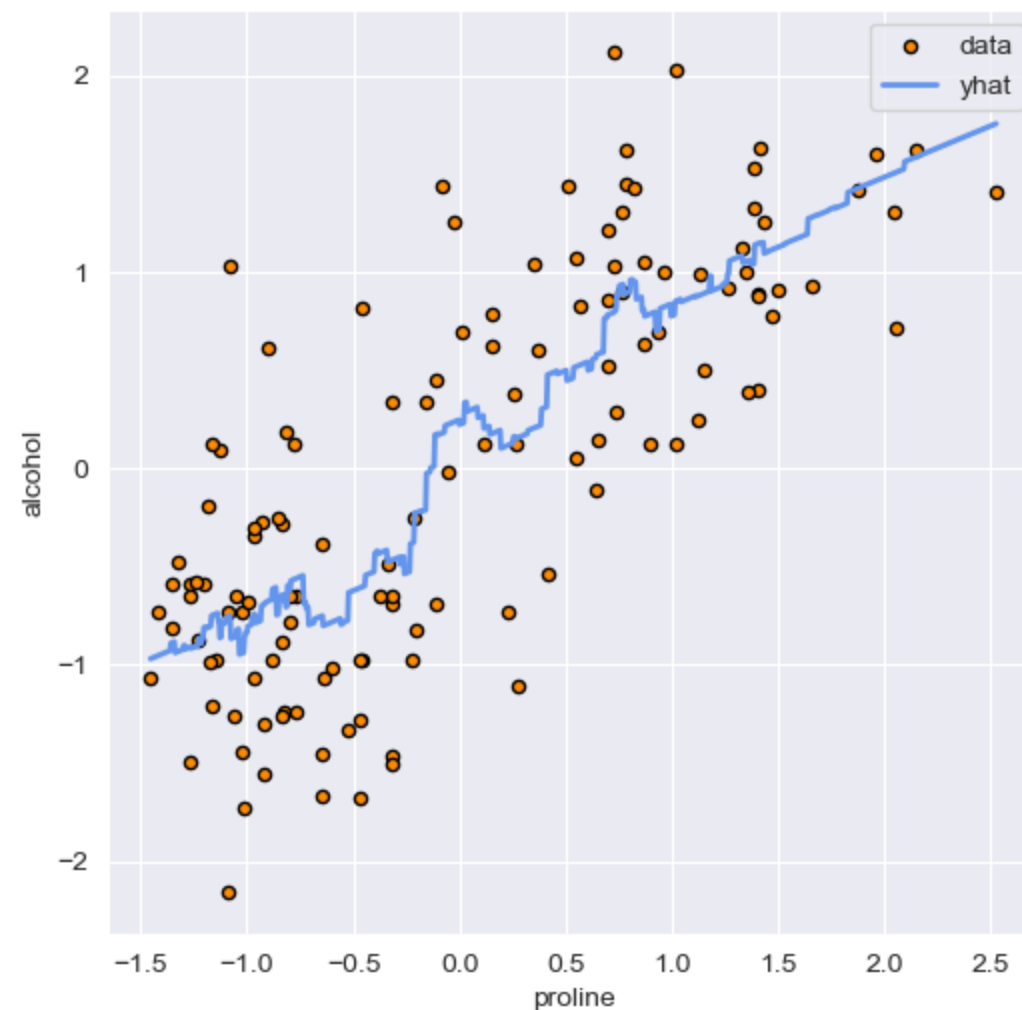
```
In [20]: 1 from sklearn.ensemble import StackingClassifier
2
3 ensemble = [('lr', LogisticRegression(max_iter=1000)),
4             ('dt', DecisionTreeClassifier(max_depth=3)),
5             ('knn', KNeighborsClassifier(n_neighbors=3))]
6
7 stackc = StackingClassifier(estimators=ensemble,
8                             final_estimator=LogisticRegression())
9 stackc.fit(X_2c_zscore, y_2c)
10
11 my_plot_decision_regions(X_2c_zscore, y_2c, stackc, figsize=(6,6))
```



Stacking for Regression

Stacking for Regression

```
In [21]: 1 from sklearn.ensemble import StackingRegressor
2
3 ensemble = [('lr', LinearRegression()),
4             ('dt', DecisionTreeRegressor(max_depth=3)),
5             ('knn', KNeighborsRegressor(n_neighbors=6))]
6
7 stackr = StackingRegressor(estimators=ensemble,
8                             final_estimator=LinearRegression())
9 stackr.fit(X_2c_zscore[['proline']], alcohol_2c_zscore)
10
11 my_plot_regression(X_2c_zscore[['proline']], alcohol_2c_zscore, stackr, figsize=(6,6))
```



Wine as Multi-Class Classification

Wine as Multi-Class Classification

```
In [22]: 1 X_mc = df_wine[['proline', 'hue']]
          2 y_mc = df_wine.target
          3
          4 X_mc_zscore = X_mc.apply(zscore,axis=0)
          5 alcohol_mc_zscore = zscore(df_wine.alcohol)
          6
          7 y_mc.value_counts().sort_index()
```

```
Out[22]: 0    59
          1    71
          2    48
          Name: target, dtype: int64
```

Multiclass and Multilabel

- **Multiclass Classification** : more than two categories/classes
 - red/green/blue, flower type, integer 0-10
- **Multilabel Classification** : can assign more than one category to an instance
 - paper topics, entities in image
- **Multiclass-Multilabel/Multitask Classification** : >1 one property with >2 one categories
 - type of fruit AND color of fruit
- **Multioutput Regression** : more than one numeric targets
 - temperature AND humidity

See sklearn docs (<https://scikit-learn.org/stable/modules/multiclass.html#>)

Sklearn Inherantly Multiclass

- `LogisticRegression(multi_class='multinomial')`
- `KNeighborsClassifier`
- `DecisionTreeClassifier`
- `RandomForestClassifier`

Sklearn Inherantly Multiclass

- `LogisticRegression(multi_class='multinomial')`
- `KNeighborsClassifier`
- `DecisionTreeClassifier`
- `RandomForestClassifier`

```
In [23]: 1 dt_mc = DecisionTreeClassifier().fit(X_mc_zscore,y_mc) # fit on multiclass
          2
          3 # generate 3 predictions
          4 y_hats = dt_mc.predict(X_mc_zscore.iloc[[82,15,166]])
          5
          6 # display target and prediction
          7 pd.DataFrame({'y':y_mc.iloc[[82,15,166]], 'y_hat':y_hats})
```

Out[23]:

	y	y_hat
82	1	1
15	0	0
166	2	2

One Vs. Rest (OvR) Classification For Multiclass

One Vs. Rest (OvR) Classification For Multiclass

- What about other models (eg Perceptron)?
- Can use any binary classifier for Multiclass classification by training multiple models:
 - model 1 : class 1 vs (class 2 and class 3)
 - model 2 : class 2 vs (class 1 and class 3)
 - model 3 : class 3 vs (class 1 and class 2)
- Then
 - Predict \hat{y} using the model with highest $P(y = \hat{y} \mid x)$, or distance from boundary, or ...

Sklearn OvR for Multiclass

- `LogisticRegression(multi_class="ovr")`
- `GradientBoostingClassifier`
- `Perceptron`

OvR For Logistic Regression

OvR For Logistic Regression

```
In [24]: 1 logr_mc = LogisticRegression(multi_class='ovr', # default
2                                     max_iter=1000)      # to avoid timeout errors
3 logr_mc.fit(X_mc_zscore,y_mc)
4 y_hats = logr_mc.predict(X_mc_zscore.iloc[[82,15,166]]) # generate 3 predictions
5 y_prob = logr_mc.predict_proba(X_mc_zscore.iloc[[82,15,166]])
6 pd.DataFrame({'y_hat':y_hats,'p_c1':y_prob[:,0],'p_c2':y_prob[:,1],'p_c3':y_prob[:,2]})\
7   .style.background_gradient(subset=['p_c1','p_c2','p_c3']).format('{:.2f}',subset=['p_c1','p_c2','p_c3'])
```

Out[24]:

	y_hat	p_c1	p_c2	p_c3
0	1	0.15	0.85	0.00
1	0	0.97	0.03	0.00
2	2	0.18	0.34	0.48

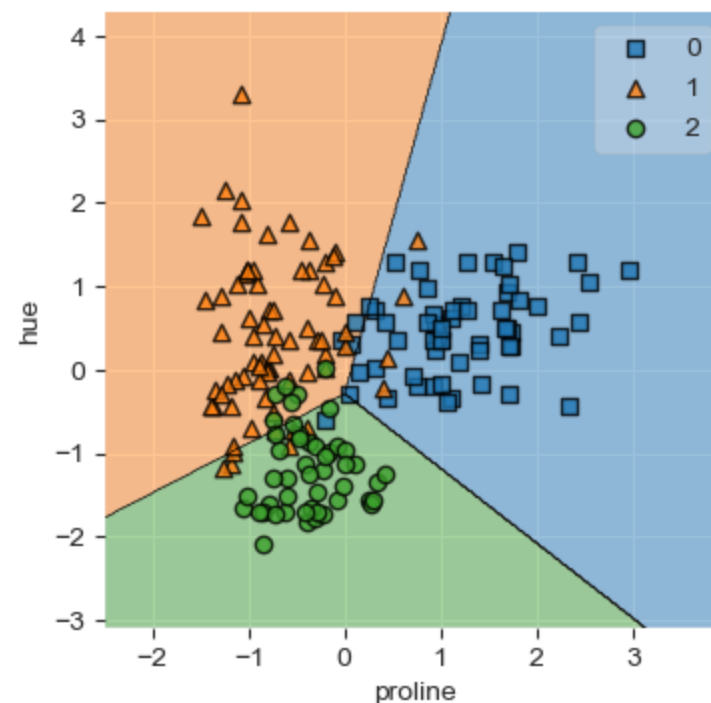
OvR For Logistic Regression

```
In [24]: 1 logr_mc = LogisticRegression(multi_class='ovr', # default
2                                     max_iter=1000)      # to avoid timeout errors
3 logr_mc.fit(X_mc_zscore,y_mc)
4 y_hats = logr_mc.predict(X_mc_zscore.iloc[[82,15,166]]) # generate 3 predictions
5 y_prob = logr_mc.predict_proba(X_mc_zscore.iloc[[82,15,166]])
6 pd.DataFrame({'y_hat':y_hats,'p_c1':y_prob[:,0],'p_c2':y_prob[:,1],'p_c3':y_prob[:,2]})\
7   .style.background_gradient(subset=['p_c1','p_c2','p_c3']).format('{:.2f}',subset=['p_c1','p_c2','p_c3'])
```

Out[24]:

	y_hat	p_c1	p_c2	p_c3
0	1	0.15	0.85	0.00
1	0	0.97	0.03	0.00
2	2	0.18	0.34	0.48

```
In [25]: 1 my_plot_decision_regions(X_mc_zscore,y_mc,logr_mc,figsize=(4,4))
```



One vs. One Classification

- Train one classifier for each pair-wise comparison of classes
- SVC

Inherantly Multilabel (aka Multioutput)

- `KNeighborsClassifier`
- `DecisionTreeClassifier`
- `MLPClassifier`
- `RandomForestClassifier`

Inherantly Multilabel (aka Multioutput)

- KNeighborsClassifier
- DecisionTreeClassifier
- MLPClassifier
- RandomForestClassifier

```
In [26]: 1 X_ml = X_mc[['proline']].iloc[-3:] # get 3 samples
          2 y_ml = np.array([[1, 0, 1], [0, 1, 1], [0, 0, 0]]) # generate 3 sets of random multi-labels for demonstration
          3
          4 dt_ml = DecisionTreeClassifier(max_depth=1).fit(X_ml,y_ml) # fit multilabel
          5 dt_ml.predict(X_ml)
```

```
Out[26]: array([[0, 0, 1],
                [0, 0, 1],
                [0, 0, 0]])
```

Sklearn `MultiOutputClassifier` meta-estimator

- fits one classifier per target (One vs. Rest)

Sklearn `MultiOutputClassifier` meta-estimator

- fits one classifier per target (One vs. Rest)

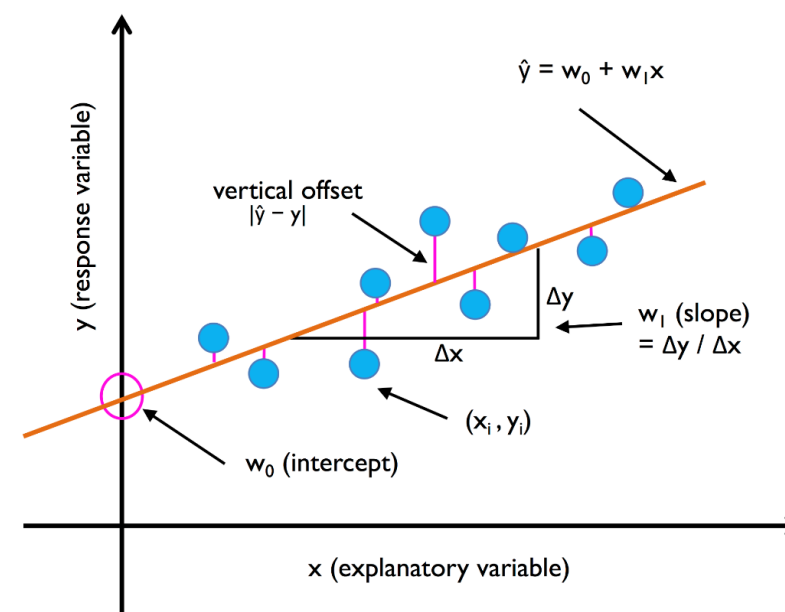
```
In [27]: 1 from sklearn.multioutput import MultiOutputClassifier
          2
          3 # wrap your classifier with MultiOutputClassifier
          4 mc_logr = MultiOutputClassifier(LogisticRegression())
          5 mc_logr.fit(X_ml,y_ml)
          6 mc_logr.predict(X_ml)
```

```
Out[27]: array([[0, 0, 1],
                [1, 1, 1],
                [0, 0, 0]])
```

Review of Models

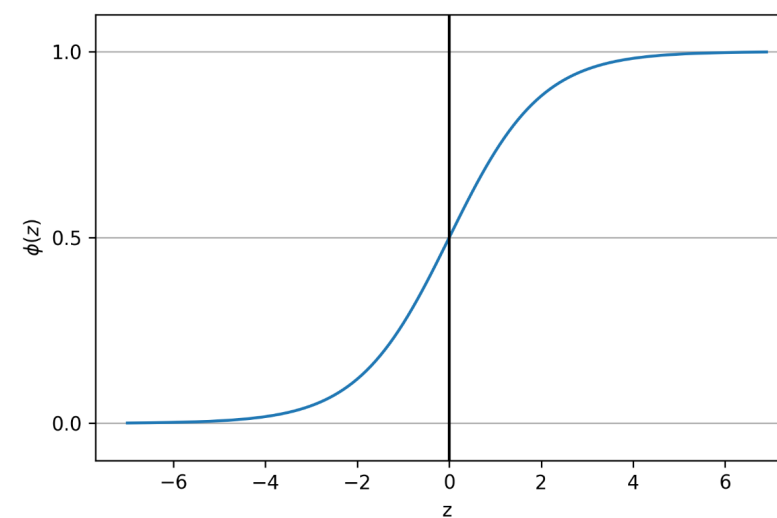
Model Review: Simple/Multiple Linear Regression

- Use for: Regression
- Pros:
 - fast to train
 - interpretable coefficients
- Cons:
 - assumes linear relationship
 - depends on removing colinear features



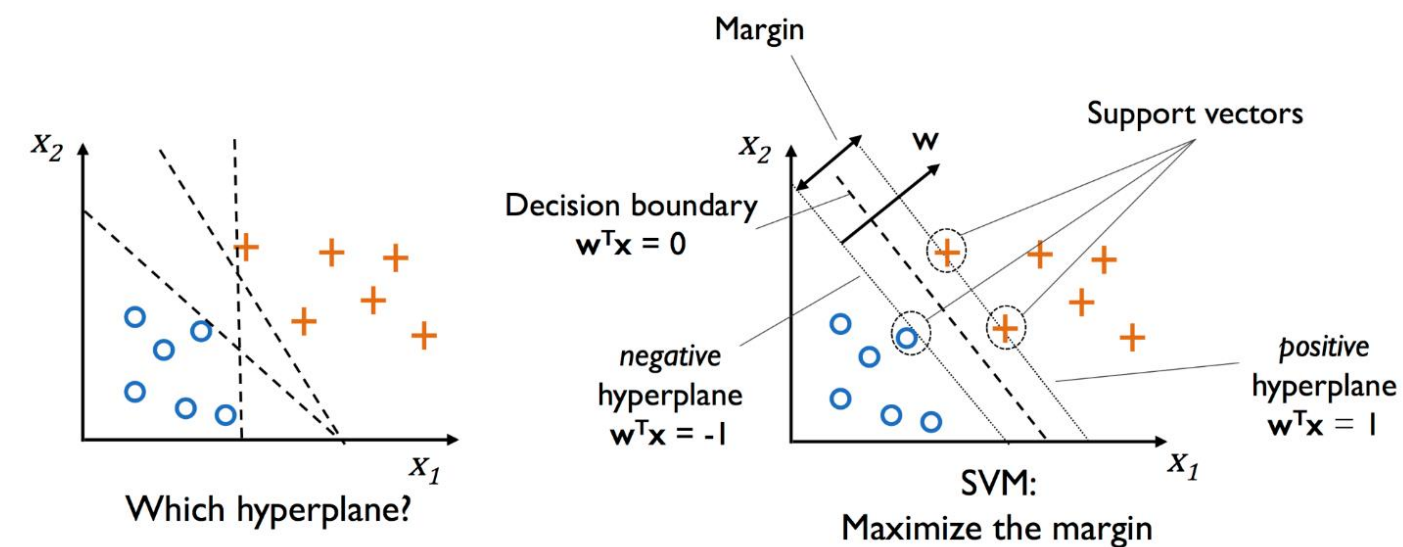
Model Review: Logistic Regression

- Use for: Classification
- Pros:
 - fast to train
 - interpretable coefficients (log odds)
- Cons:
 - assumes linear boundary
 - depends on removing colinear features



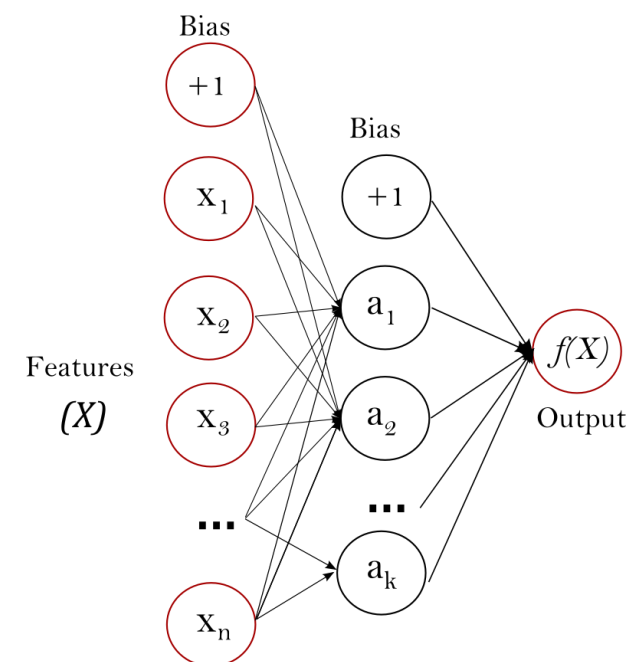
Model Review: Support Vector Machine (SVM)

- Use for: Classification and Regression
- Pros:
 - fast to evaluate
 - can use kernel trick to learn non-linear functions
- Cons:
 - slow to train
 - can fail to converge on very large datasets



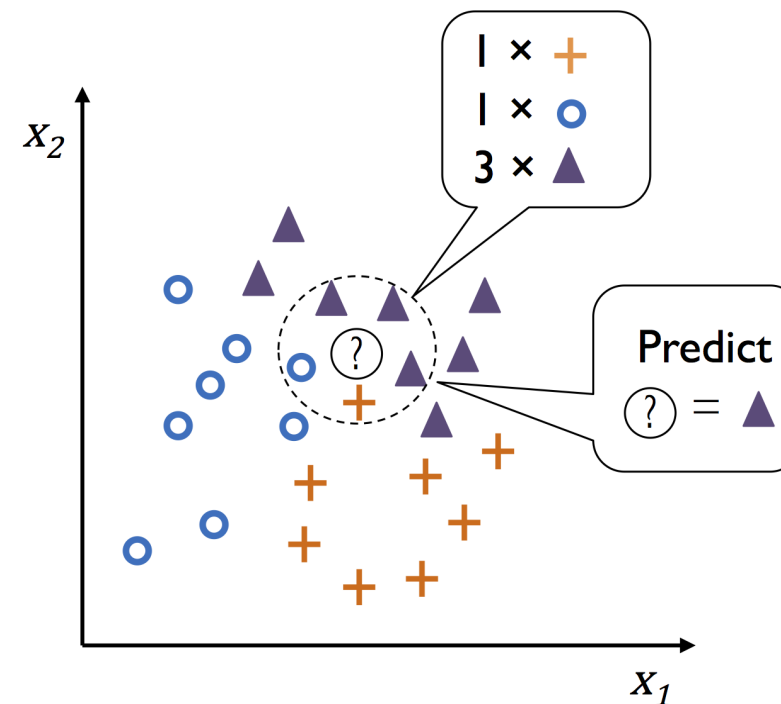
Model Review: Multi-Layer Perceptron

- Use for Classification or Regression
- Pros:
 - non-linear boundary
- Cons:
 - non-convex loss function (sensitive to initial weights)
 - sensitive to feature scaling
 - no GPU support in sklearn: use tensorflow or pytorch



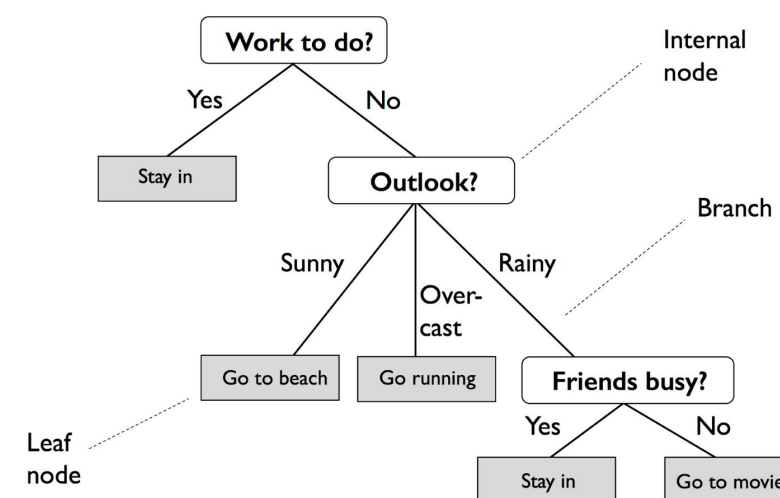
Model Review: k Nearest Neighbor (kNN)

- Use for: Classification or Regression
- Pros:
 - fast to train
 - non-linear boundary
- Cons:
 - potentially slow to predict
 - curse of dimensionality



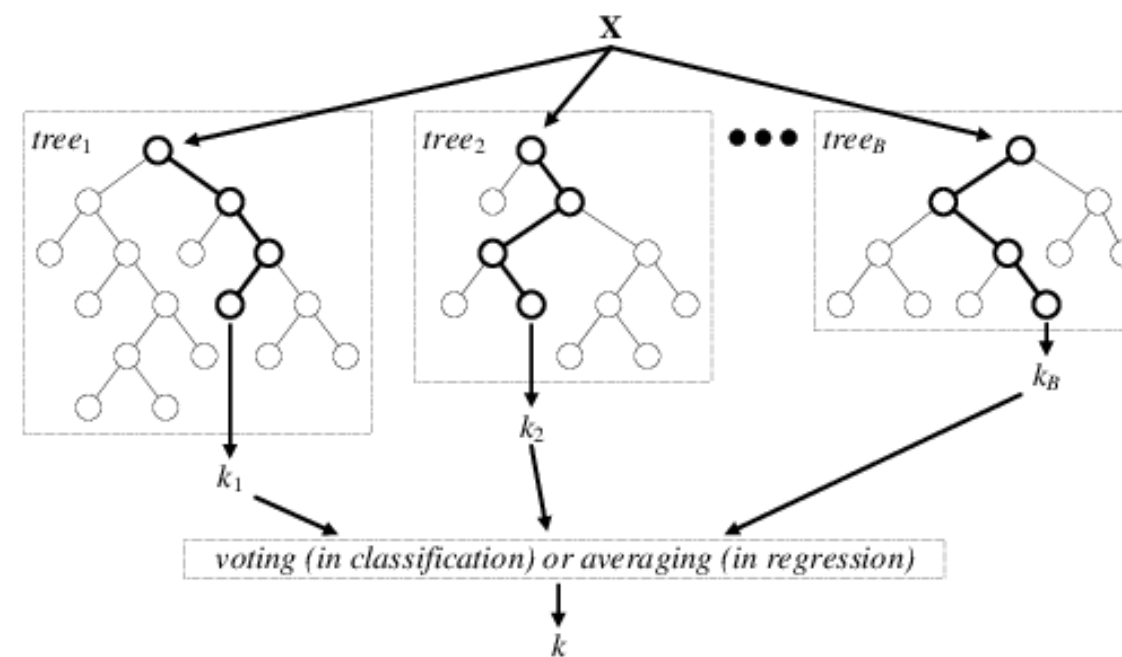
Model Review: Decision Tree

- Use for: Classification or Regression
- Pros:
 - very interpretable
 - quick to predict
 - can handle numeric and categorical variables without transformation
- Cons:
 - tendency to overfit (learn training set too well, more next class!)



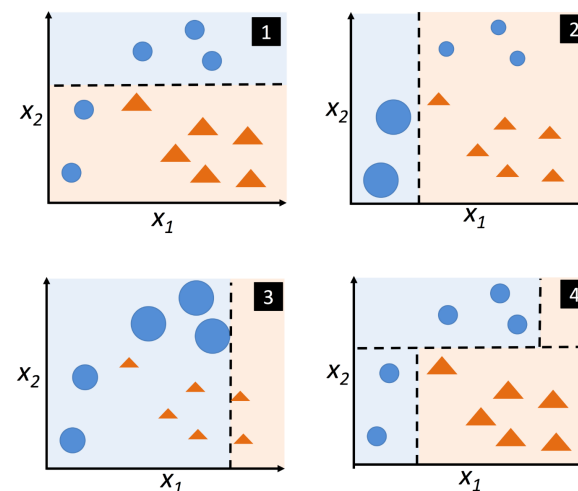
Model Review: Random Forest (Ensemble via Bagging)

- Use for: Classification or Regression
- Pros:
 - less likely to overfit than decision tree
 - quick to train (through parallelization, quick to predict)
- Cons:
 - less interpretable, though still possible



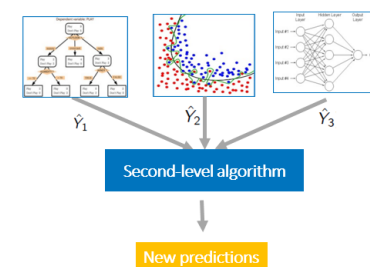
Model Review: Gradient Boosted Trees (Ensemble via Boosting)

- Use for: Classification or Regression
- Pros:
 - pays more attention to difficult decision regions
 - quick to predict
 - tends to work well on difficult tasks
- Cons:
 - slow to train (parallelization not possible)
 - less interpretable, though still possible



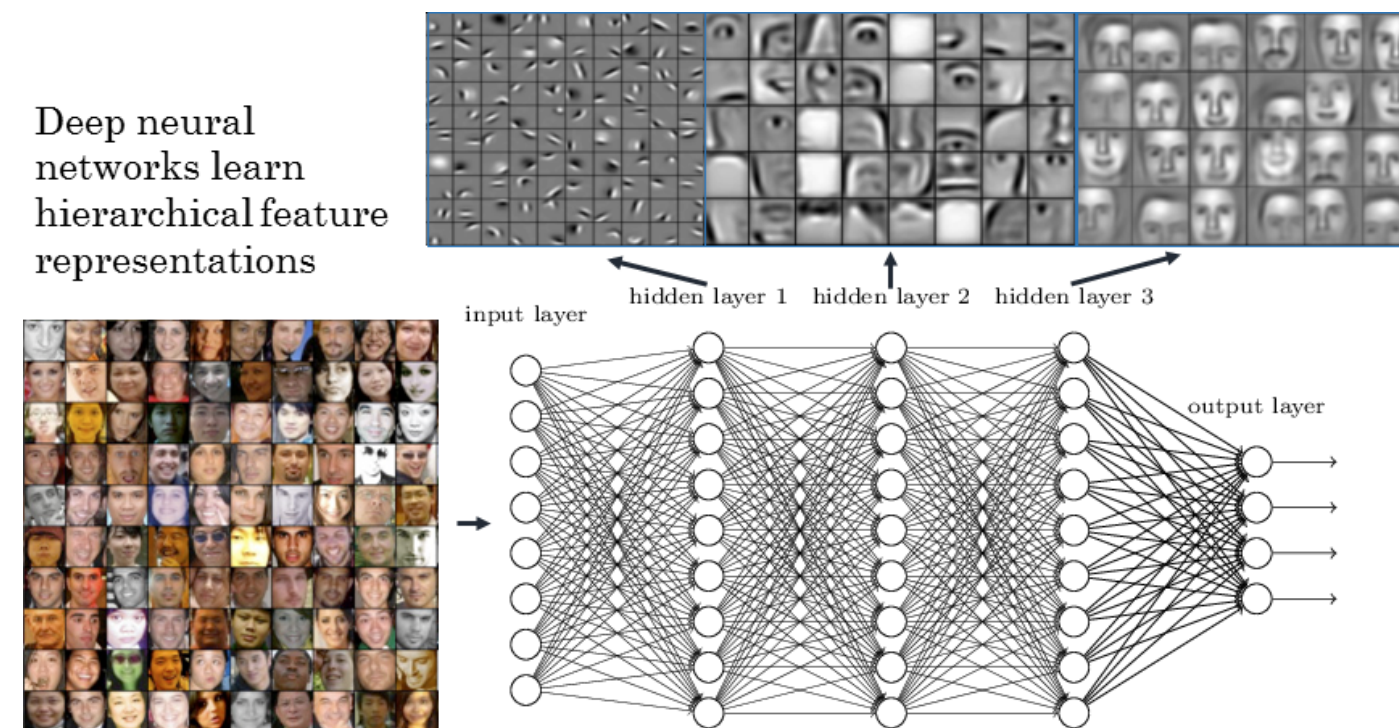
Model Review: Ensemble via Stacking

- Use for: Classification or Regression
- Pros:
 - combines benefits of multiple learning types
 - easy to implement
 - tends to win competitions
- Cons:
 - difficult to interpret
 - training/prediction time depends on component models



Neural Networks (aka Deep Learning)

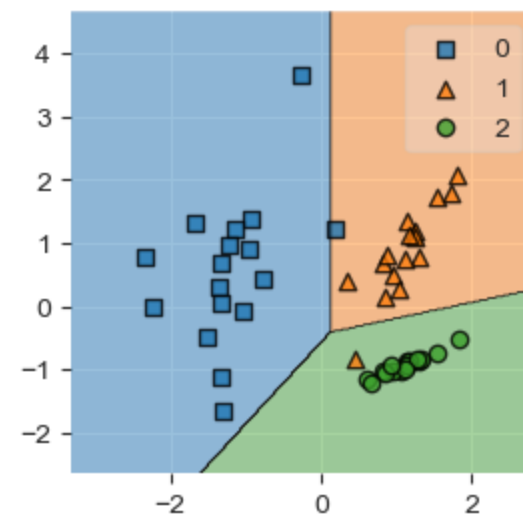
- Pros and Cons of Deep Learning
 - sensitive to initialization and structure
 - high complexity -> needs more data
 - low interpretability
 - can learn complex interactions
 - performs well on tasks involving complex signals (ex images, sound, etc)



Playing with synthetic classification datasets

Playing with synthetic classification datasets

```
In [28]: 1 from sklearn.datasets import make_classification, make_multilabel_classification
2
3 X_syn,y_syn = make_classification(n_samples=50,
4                                   n_features=2,
5                                   n_informative=2,
6                                   n_redundant=0,
7                                   n_clusters_per_class=1,
8                                   class_sep=1,
9                                   n_classes=3,
10                                  random_state=0,
11                                  )
12 fig,ax = plt.subplots(1,1,figsize=(3,3))
13 plot_decision_regions(X_syn,y_syn,LogisticRegression().fit(X_syn,y_syn));
```



Playing with synthetic classification datasets - multilabel

Playing with synthetic classification datasets - multilabel

```
In [29]: 1 X_syn_ml,y_syn_ml = make_multilabel_classification(n_samples=100,  
2                                                         n_features=2,  
3                                                         n_classes=5,  
4                                                         random_state=0  
5                                                         )  
6 print(X_syn_ml[:10])  
7 print()  
8 print(y_syn_ml[:10])
```

```
[[24. 25.]  
 [38. 15.]  
 [39. 14.]  
 [23. 20.]  
 [26. 29.]  
 [30. 16.]  
 [22. 30.]  
 [25. 22.]  
 [29. 12.]  
 [25. 21.]]
```

```
[[0 0 0 0 0]  
 [0 0 1 0 1]  
 [0 1 1 1 1]  
 [1 1 1 1 1]  
 [1 0 0 1 0]  
 [0 1 0 1 1]  
 [1 0 0 0 0]  
 [1 1 1 1 0]  
 [0 0 0 0 0]  
 [0 0 1 1 0]]
```


Questions?