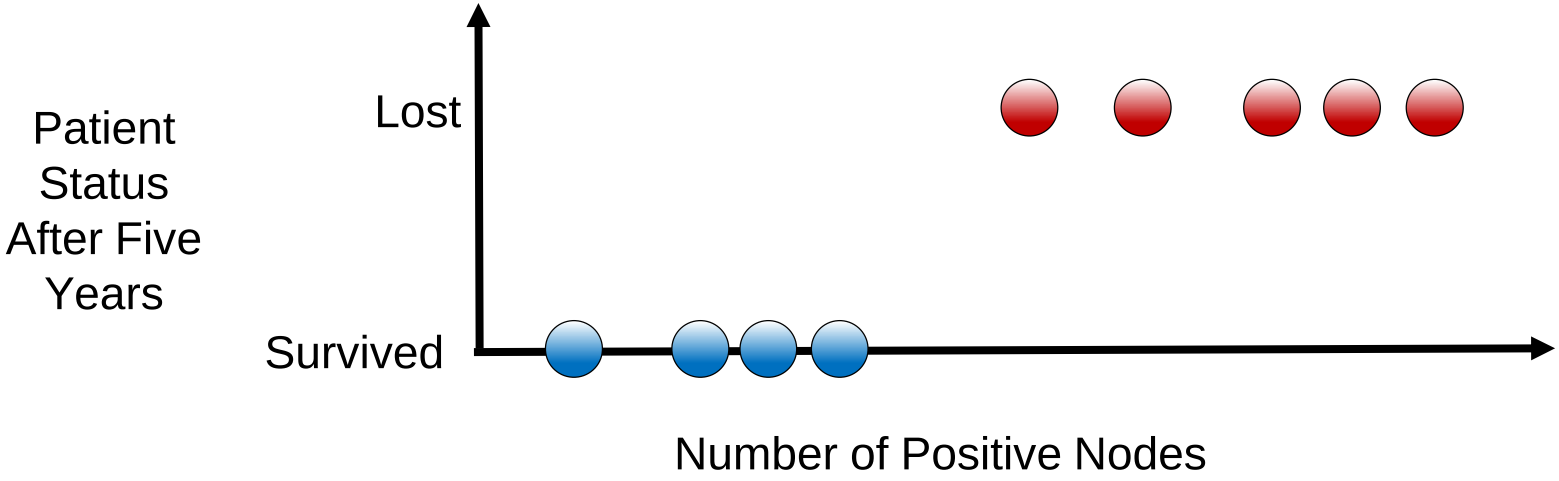
Chapter 5

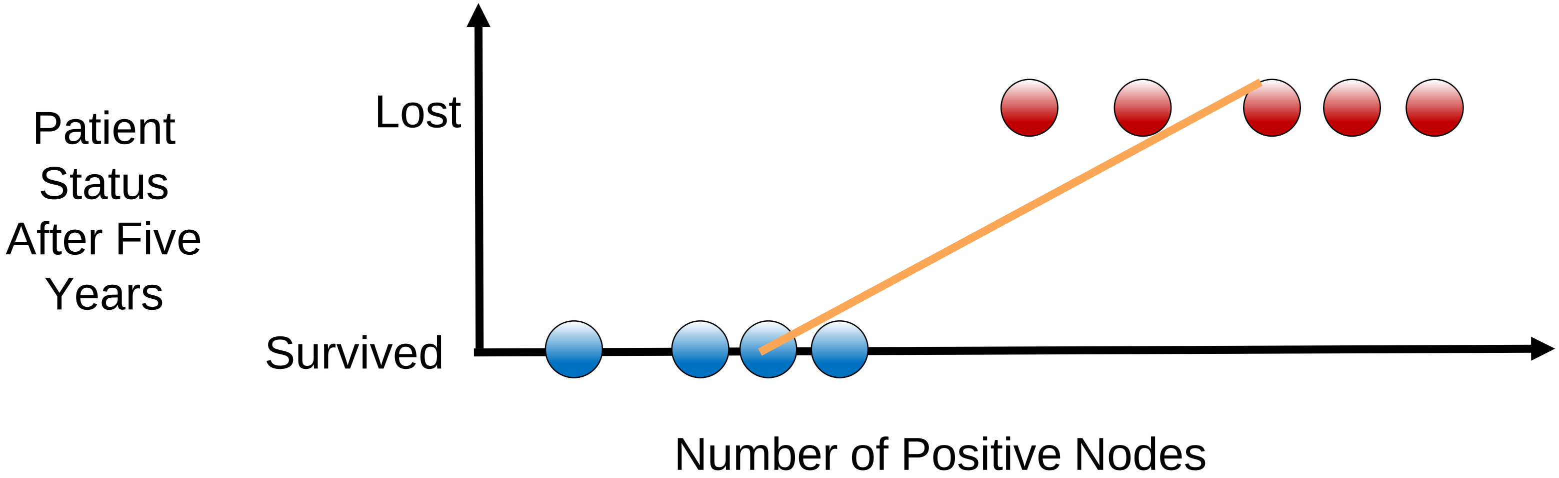
Logistic Regression
Classification Error Metrics

Logistic Regression

Introduction to Logistic Regression

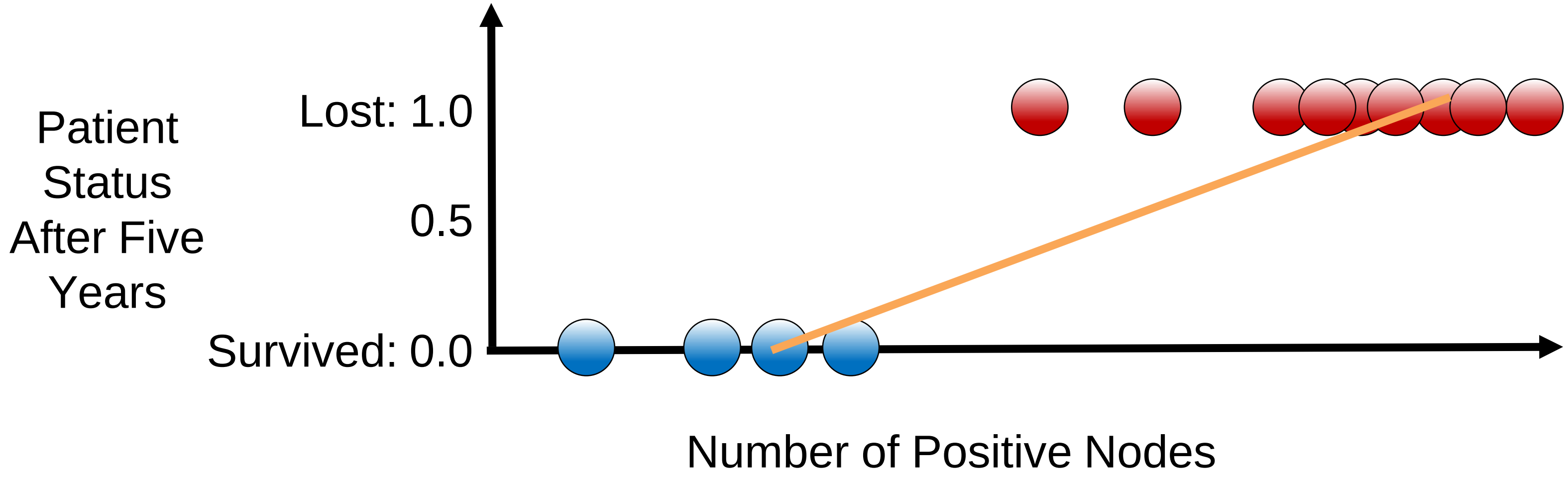


Linear Regression for Classification?



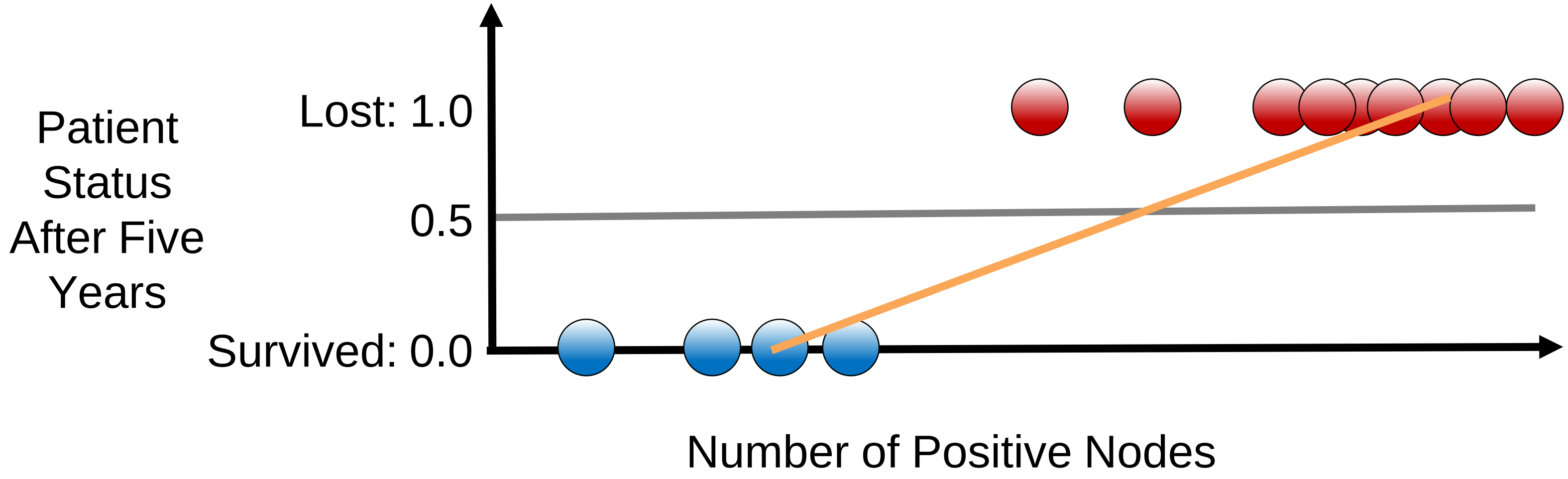
$$y_\beta(x) = \beta_0 + \beta_1 x + \varepsilon$$

Linear Regression for Classification?



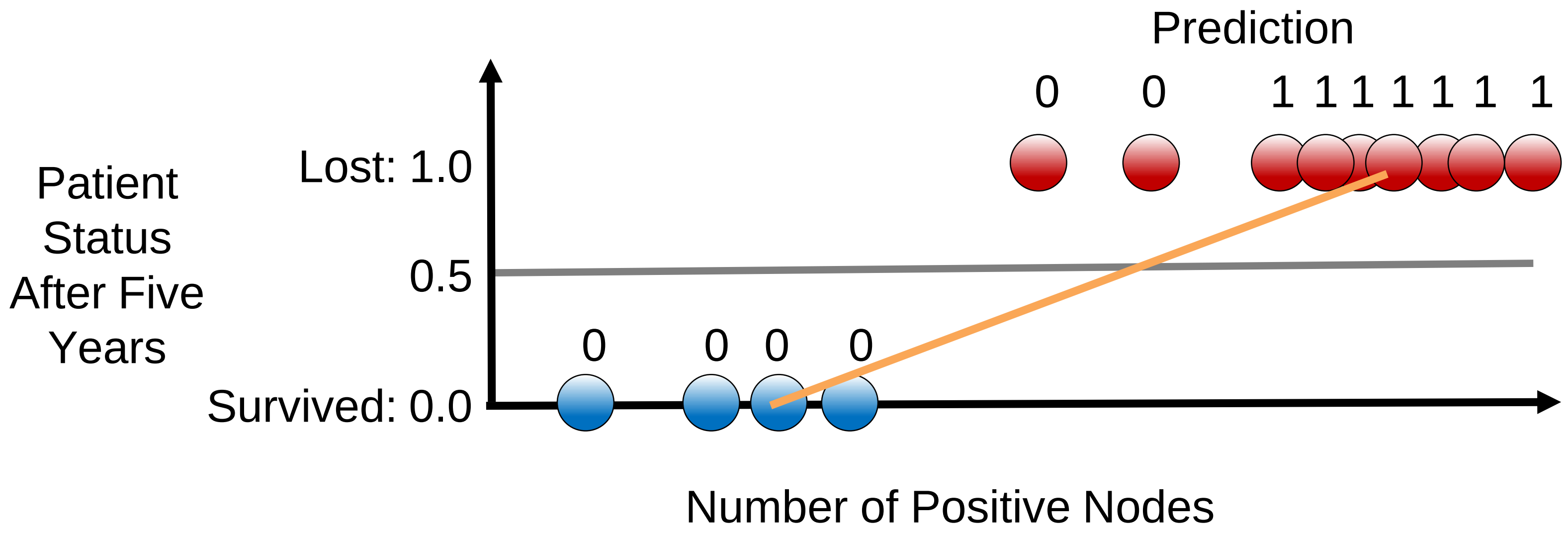
$$y_{\beta}(x) = \beta_0 + \beta_1 x + \varepsilon$$

Linear Regression for Classification?



If model result > 0.5: predict lost
If model result < 0.5: predict survived

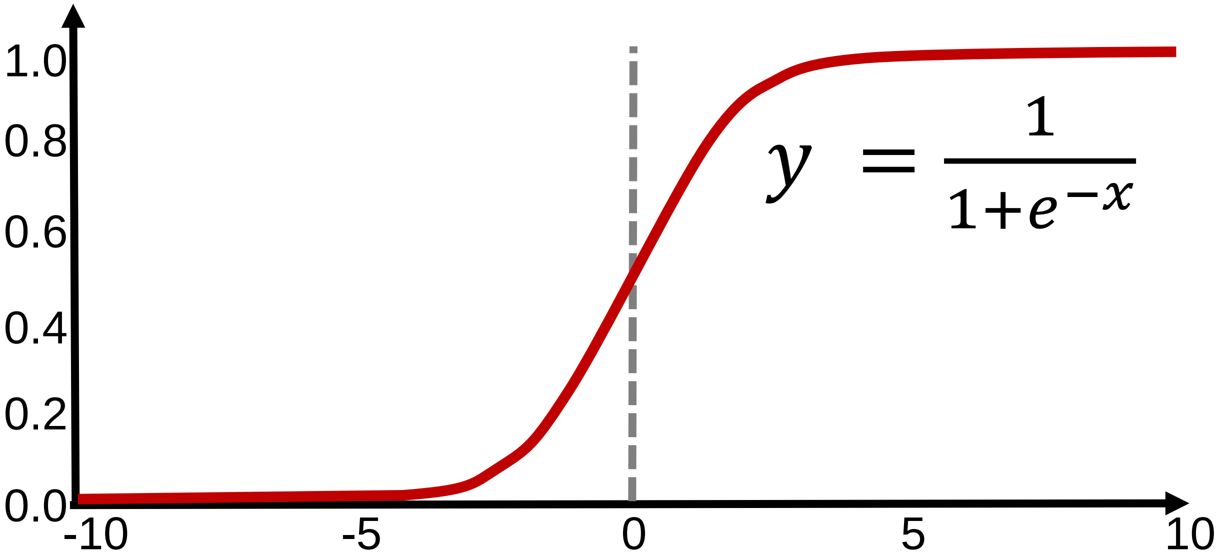
Linear Regression for Classification?



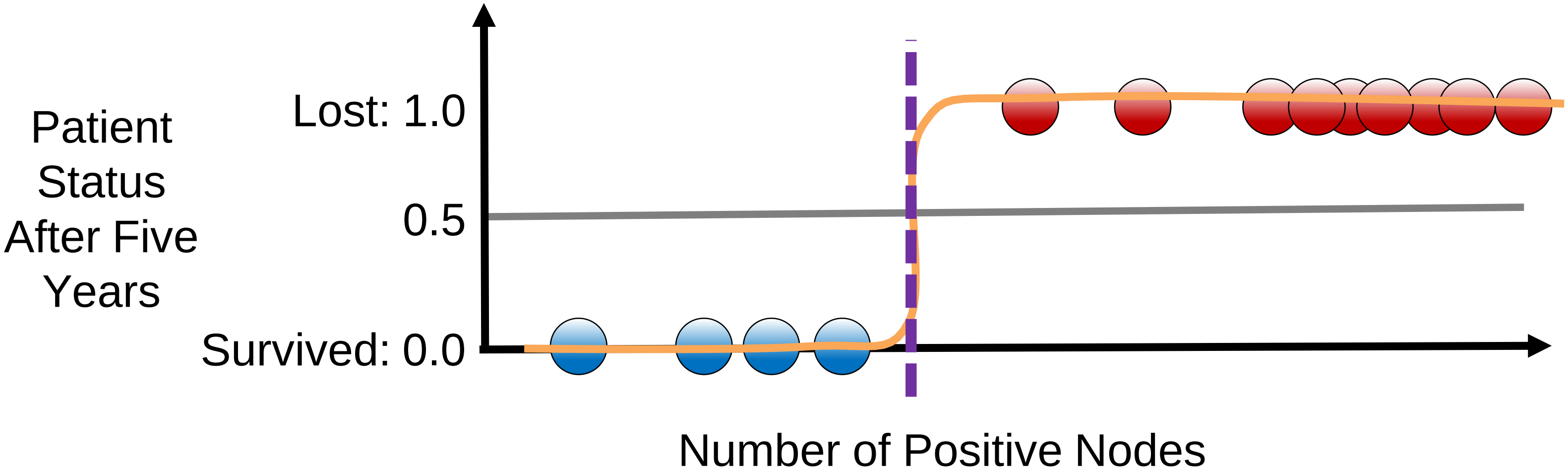
If model result > 0.5: predict lost

If model result < 0.5: predict survived

What is this Function?

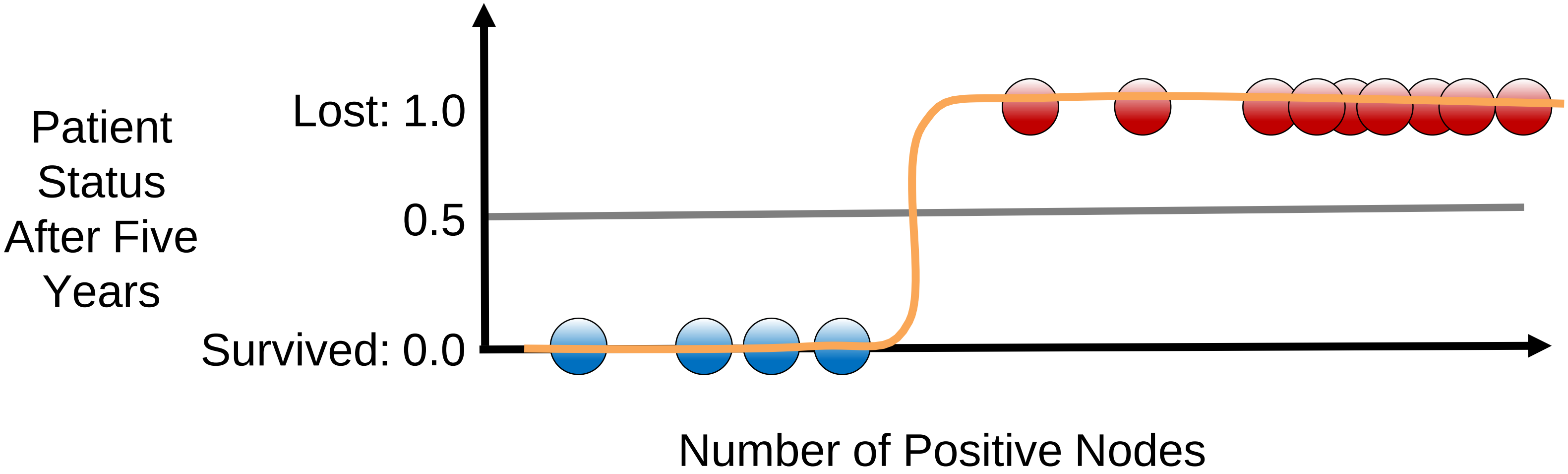


The Decision Boundary



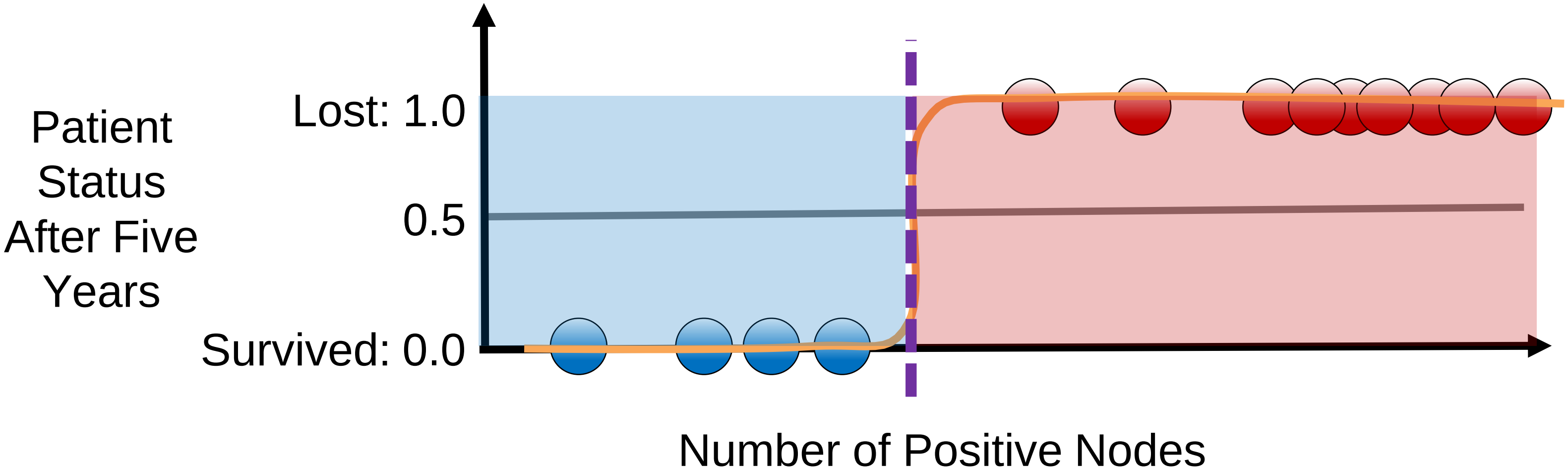
$$y_{\beta}(x) = \frac{1}{1+e^{-(\beta_0 + \beta_1 x + \varepsilon\ )}}$$

Logistic Regression



$$y_\beta(x) = \frac{1}{1+e^{-(\beta_0 + \beta_1 x + \varepsilon)}}$$

The Decision Boundary



$$y_\beta(x) = \frac{1}{1+e^{-(\beta_0+\ \beta_1 x\ +\ \varepsilon\ )}}$$

Relationship of Logistic to Linear Regression

Logistic Function

$$P(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x + \varepsilon\ )}}$$

Relationship of Logistic to Linear Regression

Logistic Function

$$P(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x + \varepsilon\ )}}$$

$$P(x) = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}}$$

Relationship of Logistic to Linear Regression

Logistic Function

$$P(x) = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}}$$

Relationship of Logistic to Linear Regression

Logistic Function

$$P(x) = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}}$$

$$\Downarrow$$

Odds Ratio

$$\frac{P(x)}{1 - P(x)} = e^{(\beta_0 + \beta_1 x)}$$

Relationship of Logistic to Linear Regression

Logistic Function

$$P(x) = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}}$$

Log Odds

$$log\left[\frac{P(x)}{1 - P(x)}\right] = \beta_0 + \beta_1 x$$

Relationship of Logistic to Linear Regression

Logistic Function

$$P(x) = \frac{e^{(\beta_0 + \beta_1 x)}}{1+e^{(\beta_0 + \beta_1 x)}}$$

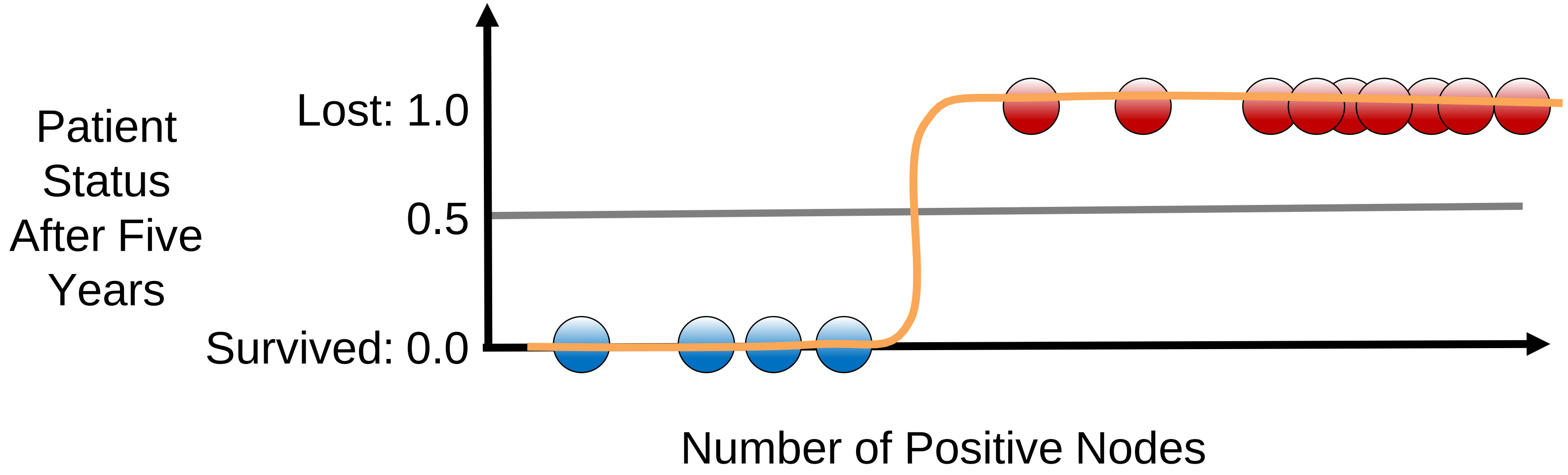
Log Odds

$$log\left[\frac{P(x)}{1-P(x)}\right] = \boxed{\beta_0 + \beta_1 x}$$

Classification with Logistic Regression

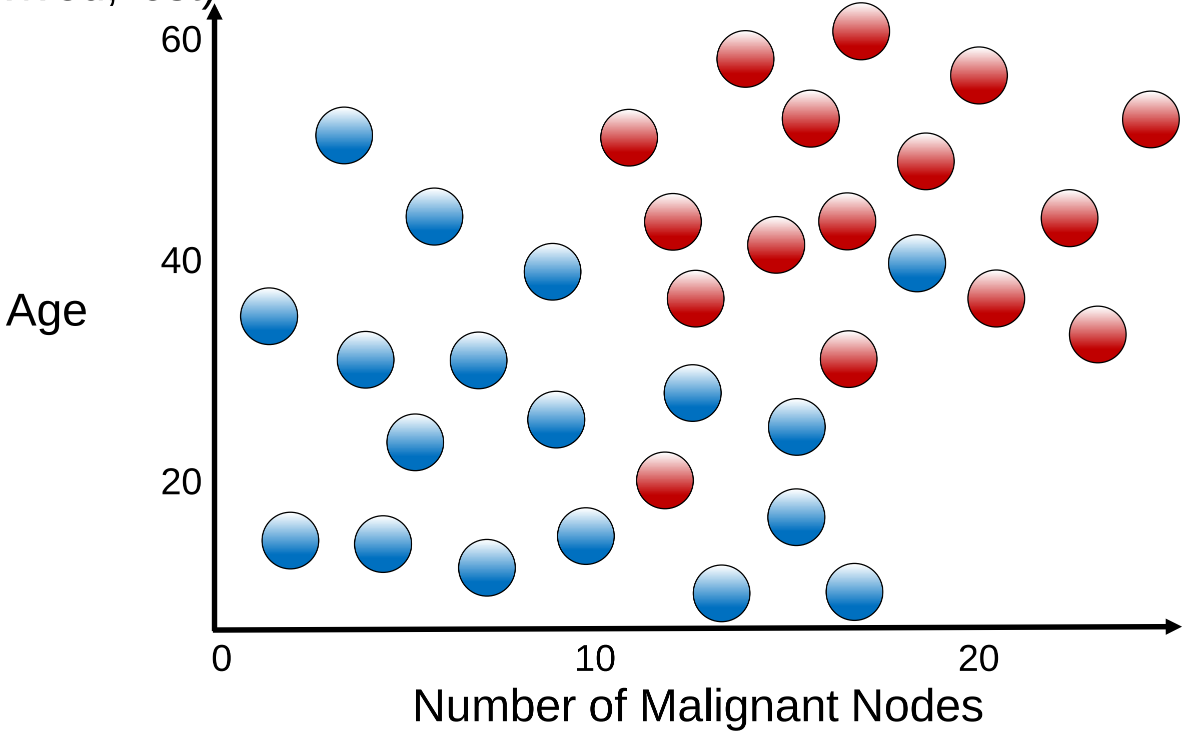
One feature (nodes)

Two labels (survived, lost)



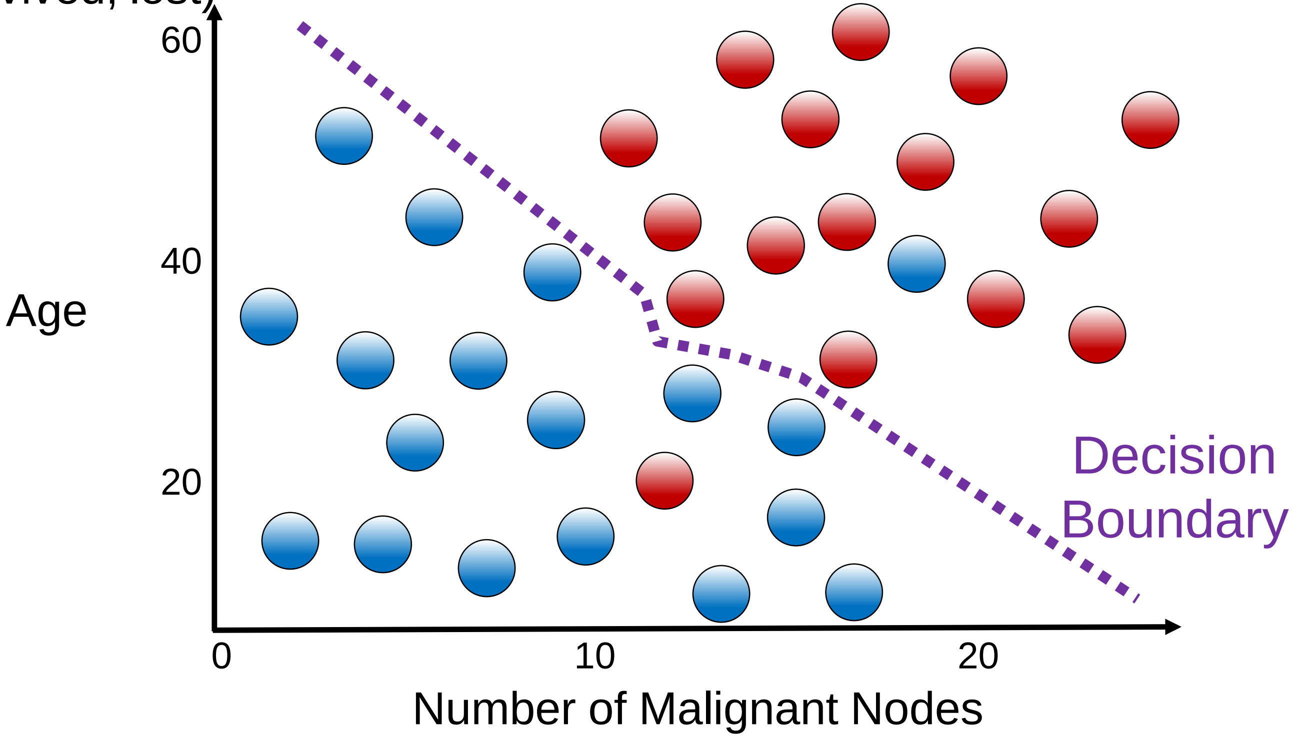
Classification with Logistic Regression

Two features (nodes, age)
Two labels (survived, lost)



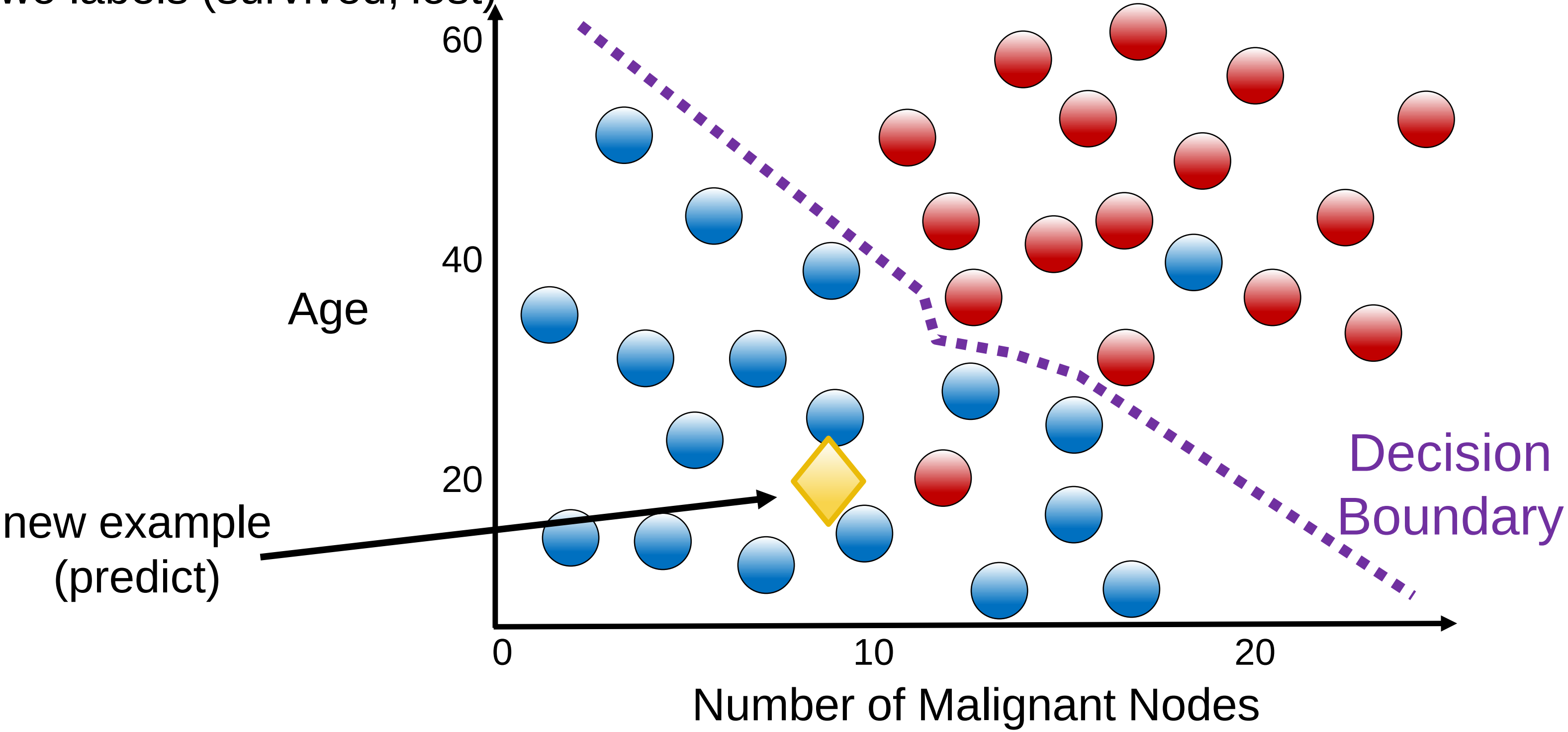
Classification with Logistic Regression

Two features (nodes, age)
Two labels (survived, lost)



Classification with Logistic Regression

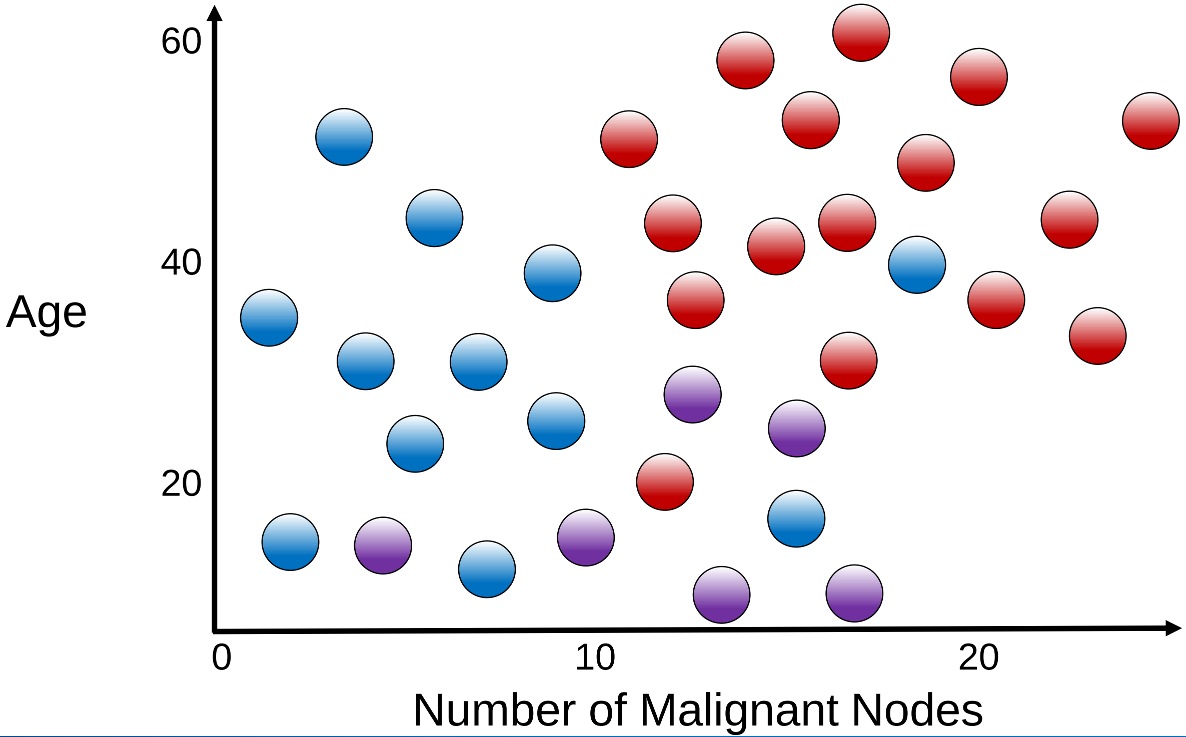
Two features (nodes, age)
Two labels (survived, lost)



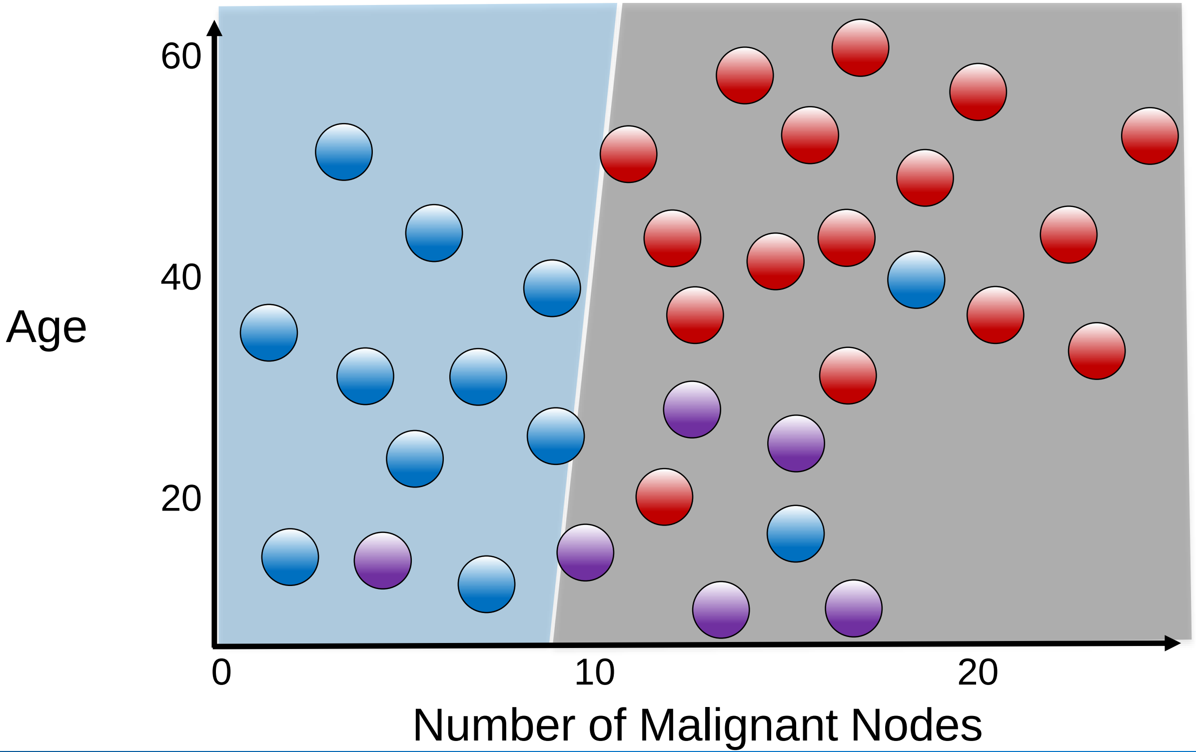
Multiclass Classification with Logistic Regression

Two features (nodes, age)

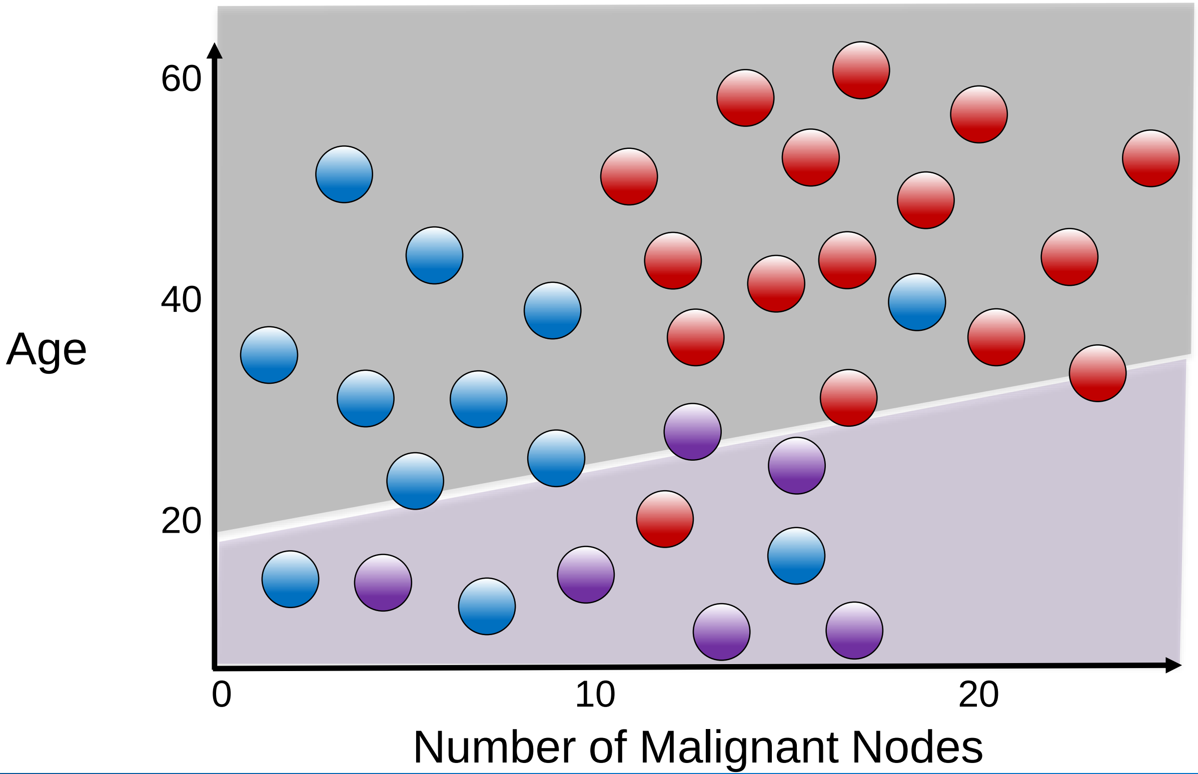
Three labels (survived, complications, lost)



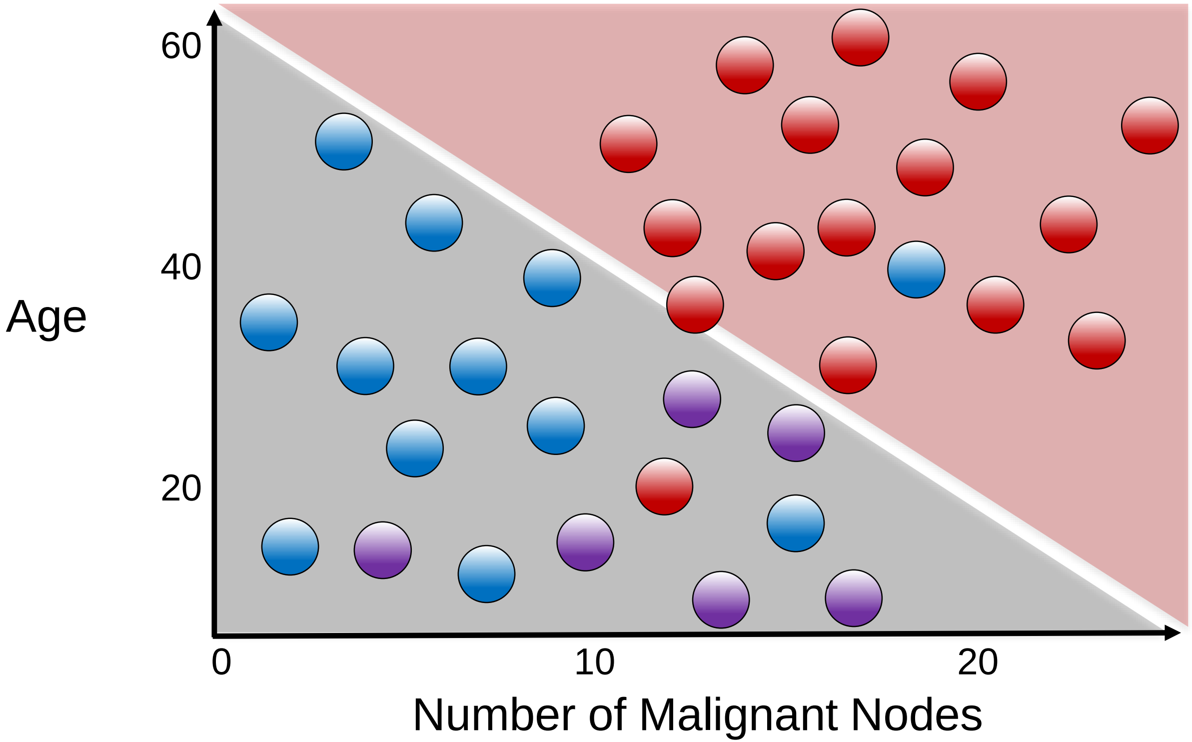
One vs All: Survived vs All



One vs All: Complications vs All

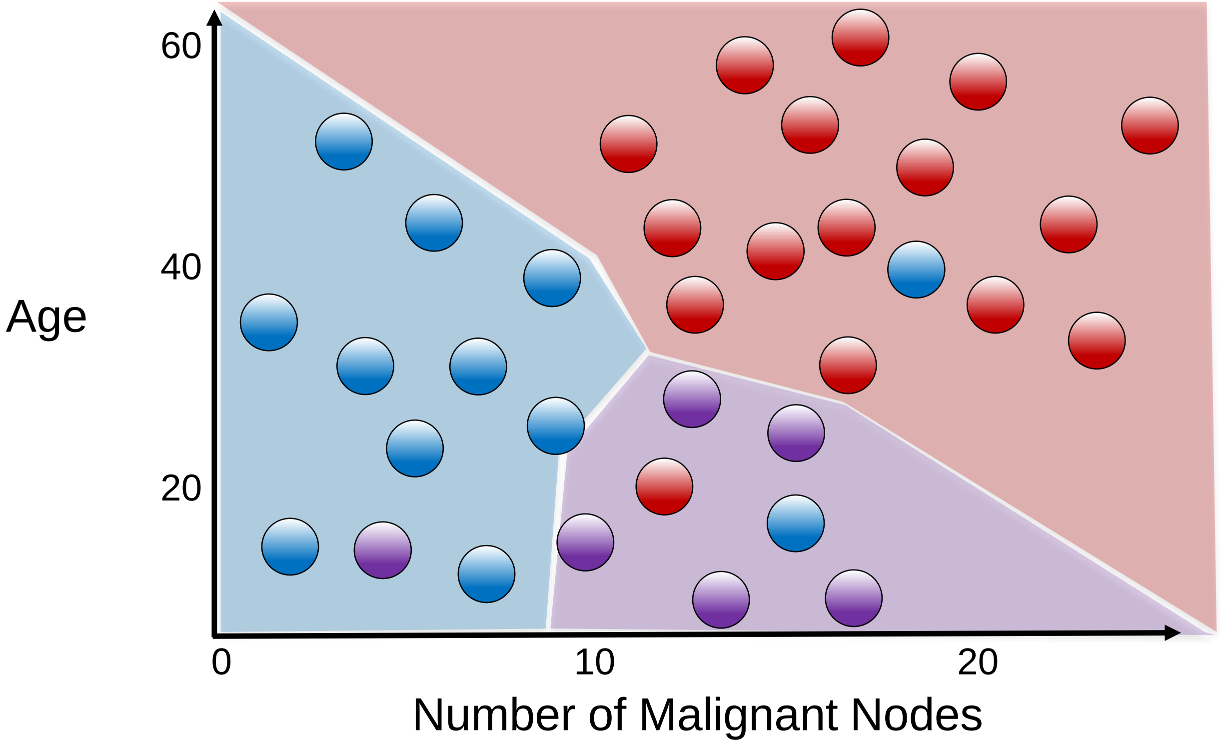


One vs All: Loss vs All



Multiclass Decision Boundary

Assign most probable class to each region



Logistic Regression: The Syntax

Import the class containing the classification method

```
from sklearn.linear_model import LogisticRegression
```

Logistic Regression: The Syntax

Import the class containing the classification method

```
from sklearn.linear_model import LogisticRegression
```

Create an instance of the class

```
LR = LogisticRegression(penalty='l2', c=10.0)
```

Logistic Regression: The Syntax

Import the class containing the classification method

```
from sklearn.linear_model import LogisticRegression
```

Create an instance of the class

```
LR = LogisticRegression(penalty='l2', c=10.0)
```

regularization parameters

Logistic Regression: The Syntax

Import the class containing the classification method

```
from sklearn.linear_model import LogisticRegression
```

Create an instance of the class

```
LR = LogisticRegression(penalty='l2', c=10.0)
```

Fit the instance on the data and then predict the expected value

```
LR = LR.fit(X_train, y_train)
y_predict = LR.predict(X_test)
```

Logistic Regression: The Syntax

Import the class containing the classification method

```
from sklearn.linear_model import LogisticRegression
```

Create an instance of the class

```
LR = LogisticRegression(penalty='l2', c=10.0)
```

Fit the instance on the data and then predict the expected value

```
LR = LR.fit(X_train, y_train)
y_predict = LR.predict(X_test)
```

Tune regularization parameters with cross-validation: LogisticRegressionCV.

Classification Error Metrics

Choosing the Right Error Measurement

- You are asked to build a classifier for leukemia
- **Training data**: 1% patients with leukemia, 99% healthy
- **Measure accuracy**: total % of predictions that are correct

Choosing the Right Error Measurement

- You are asked to build a classifier for leukemia
- **Training data**: 1% patients with leukemia, 99% healthy
- **Measure accuracy**: total % of predictions that are correct
- Build a simple model that always predicts "healthy"
- Accuracy will be 99%...

Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Confusion Matrix

	Predicted Positive	Predicted Negative	
Actual Positive	True Positive (TP)	False Negative (FN)	Type II Error
Actual Negative	False Positive (FP)	True Negative (TN)	

Type I Error

Accuracy: Predicting Correctly

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

Recall: Identifying All Positive Instances

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

$$\text{Recall or Sensitivity} = \frac{TP}{TP + FN}$$

Precision: Identifying Only Positive Instances

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

$$\text{Precision} = \frac{TP}{TP + FP}$$

Specificity: Avoiding False Alarms

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

$$\text{Specificity} = \frac{TN}{FP + TN}$$

Error Measurements

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

Error Measurements

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

$$\text{Recall or Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Specificity} = \frac{TN}{FP + TN}$$

Error Measurements

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

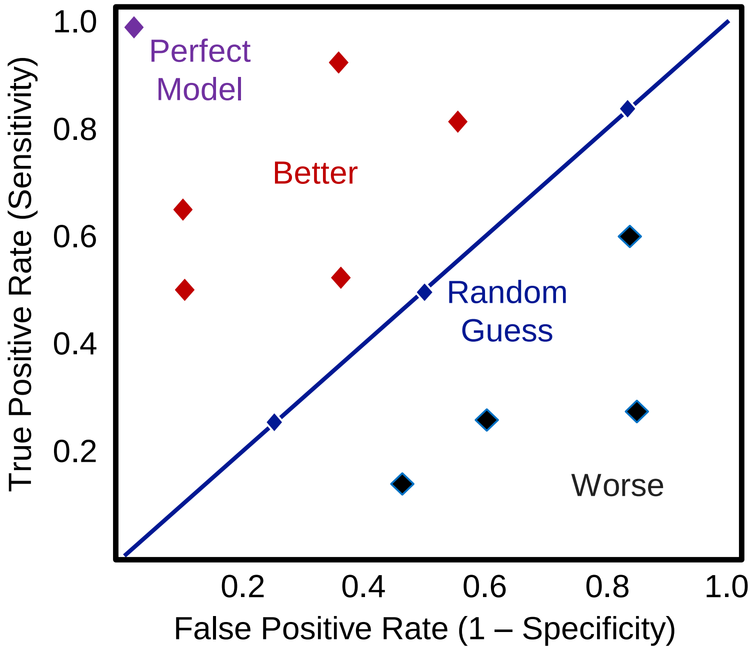
$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall or Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{FP + TN}$$

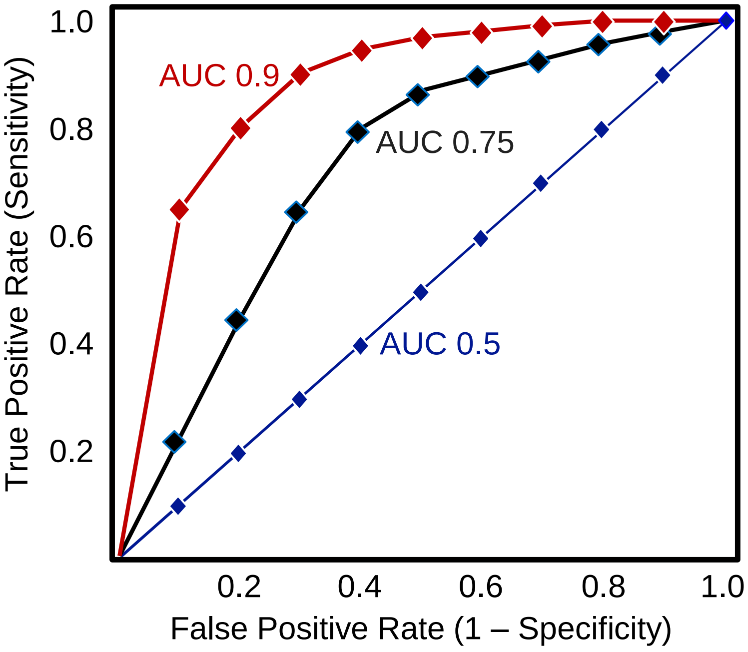
$$F1 = 2\,\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Receiver Operating Characteristic (ROC)



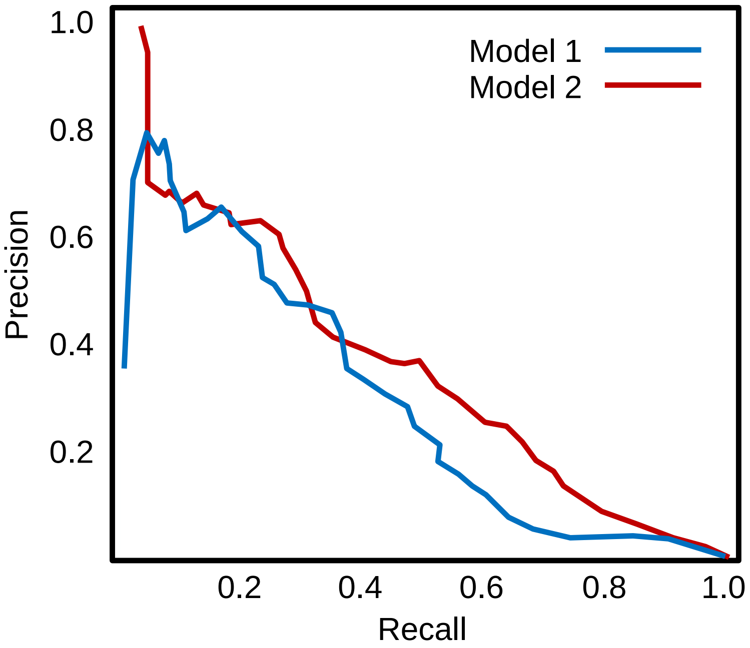
Evaluation of model at all possible thresholds

Area Under Curve (AUC)



Measures total area under ROC curve

Precision Recall Curve (PR Curve)



Measures trade-off between precision and recall

Multiple Class Error Metrics

	Predicted Class 1	Predicted Class 2	Predicted Class 3
Actual Class 1	TP1		
Actual Class 2		TP2	
Actual Class 3			TP3

Multiple Class Error Metrics

	Predicted Class 1	Predicted Class 2	Predicted Class 3
Actual Class 1	TP1		
Actual Class 2		TP2	
Actual Class 3			TP3

$$\text{Accuracy} = \frac{TP1 + TP2 + TP3}{\text{Total}}$$

Multiple Class Error Metrics

	Predicted Class 1	Predicted Class 2	Predicted Class 3
Actual Class 1	TP1		
Actual Class 2		TP2	
Actual Class 3			TP3

$$\text{Accuracy} = \frac{TP1 + TP2 + TP3}{\text{Total}}$$

Most multi-class error metrics are similar to binary versions—just expand elements as a sum

Classification Error Metrics: The Syntax

Import the desired error function

```
from sklearn.metrics import accuracy_score
```

Classification Error Metrics: The Syntax

Import the desired error function

```
from sklearn.metrics import accuracy_score
```

Calculate the error on the test and predicted data sets

```
accuracy_value = accuracy_score(y_test, y_pred)
```

Classification Error Metrics: The Syntax

Import the desired error function

```
from sklearn.metrics import accuracy_score
```

Calculate the error on the test and predicted data sets

```
accuracy_value = accuracy_score(y_test, y_pred)
```

Lots of other error metrics and diagnostic tools:

```
from sklearn.metrics import precision_score, recall_score,
                           f1_score, roc_auc_score,
                   confusion_matrix, roc_curve,
                   precision_recall_curve
```

Legal Notices and Disclaimers