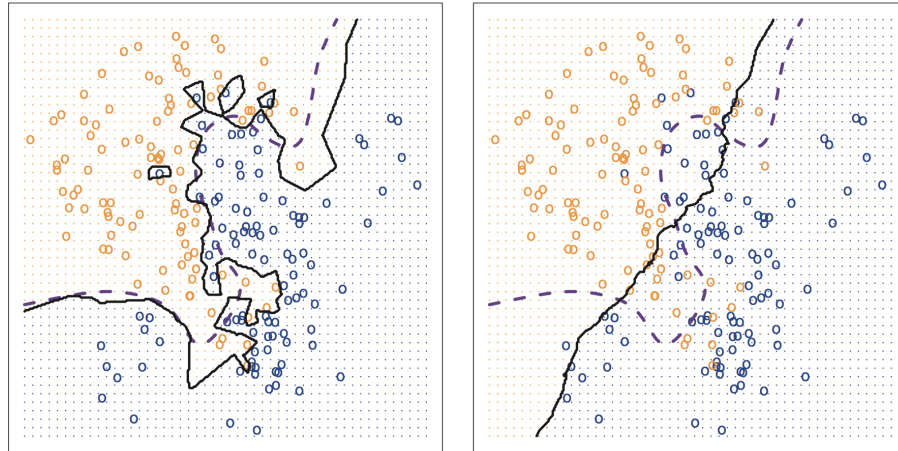


K-nearest neighbors and Featurization

(find and work with a partner)

- In the two images below, the solid lines represent the KNN *decision boundaries* (the dashed lines represent “ideal” decision boundaries – we will discuss this later). In each case, what is your guess for the value of K that was used?



- In the scenarios above, which would you describe as *underfitting*? As *overfitting*?
- Say I am trying to classify pictures of food and I have one *feature* called “shape”. The *feature values* for shape are circle, square, and triangle. However, for my ML algorithm I need numerical values. How would you *featurize* this data?
- Using your response from the previous question, what would the *feature vector* become for \mathbf{x}_1 ?

Day	Outlook	Temperature	Humidity	Wind	PlayTennis (y)
\mathbf{x}_1	Sunny	Hot	High	Weak	No
\mathbf{x}_2	Sunny	Hot	High	Strong	No
\mathbf{x}_3	Overcast	Hot	High	Weak	Yes
\mathbf{x}_4	Rain	Mild	High	Weak	Yes
\mathbf{x}_5	Rain	Cool	Normal	Weak	Yes
\mathbf{x}_6	Rain	Cool	Normal	Strong	No
\mathbf{x}_7	Overcast	Cool	Normal	Strong	Yes
\mathbf{x}_8	Sunny	Mild	High	Weak	No
\mathbf{x}_9	Sunny	Cool	Normal	Weak	Yes
\mathbf{x}_{10}	Rain	Mild	Normal	Weak	Yes

Data from Machine Learning by Tom Mitchell (Table 3.2)

- (outside class)* If I have n training examples and m testing examples, each with p features, what is the runtime of K -nearest neighbors?