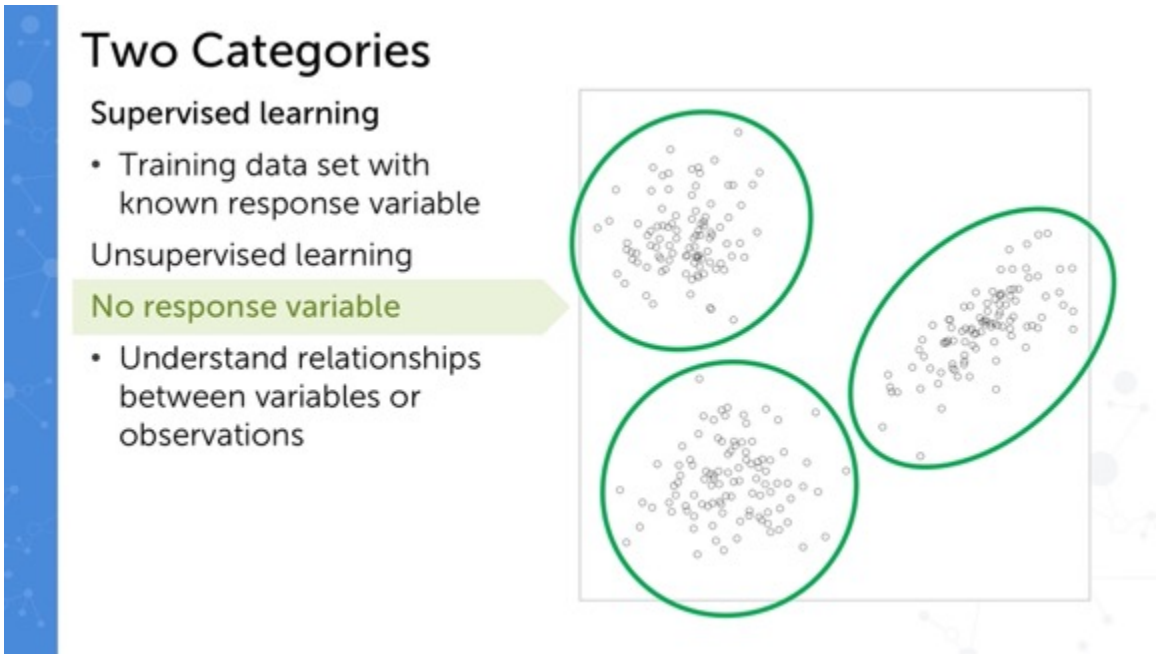
Learning Objectives

By the end of this lesson, you will be able to:

- Explain the 2 main types of data mining approaches.
- Explain the 2 types of supervised learning.
- Explain 2 reasons why the method that best fits the data is not always the best method to use.
- Estimate the error on new data points for regression and classification problems.
- Explain why interpreting a model is important, and some key questions to start the interpretation.

Methods of data mining fall into two main categories. The first is called supervised learning. This is where you have a training dataset with a known response variable, or y variable. And you want to estimate, or predict, the values of that response variable in another dataset called the test set, or validation set. Supervised learning can also be used to do inference, if you want to understand the model that describes the relationship between the predictor variables, x variables, and the response variable.

Some examples of supervised learning approaches include linear regression, logistic regression, linear discriminant analysis, and decision trees.



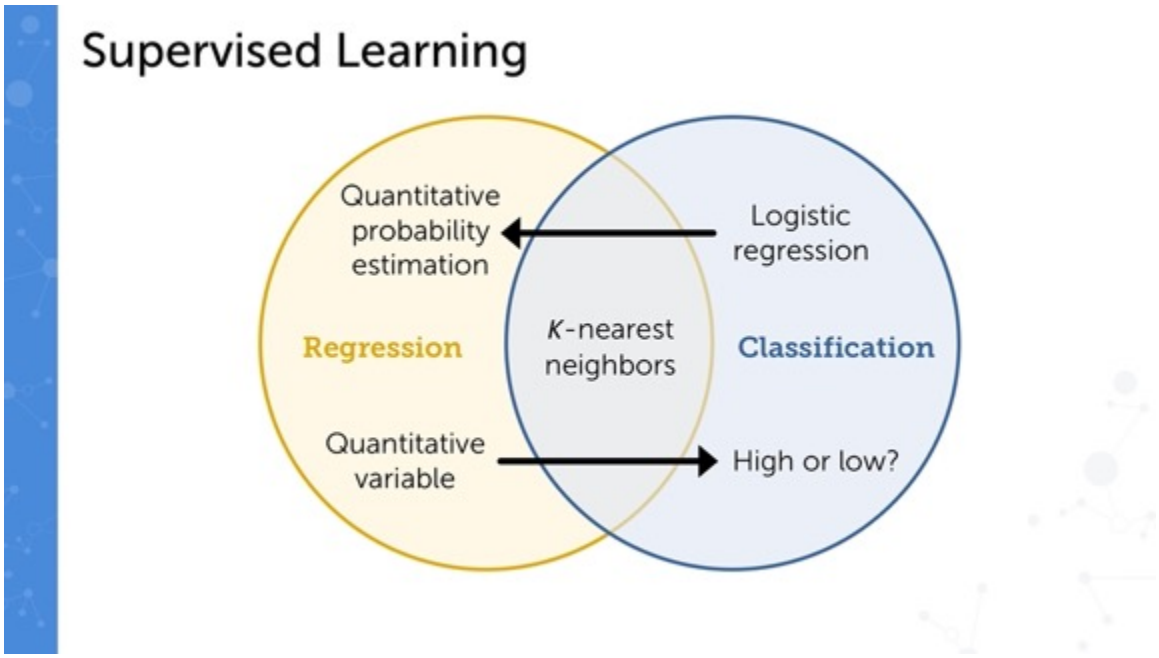
The other main category is unsupervised learning. In unsupervised learning, you have no response variable. Instead, your main goal is to understand any relationships between the variables or between the observations. For example, suppose we have two variables represented on the x and y axes here. We're not thinking about the y variable as being a response, but we want to understand its relationship with the x variable. And in this case, we could use cluster analysis to identify three main clusters in the data.

Some other examples of unsupervised learning and data mining are principal components analysis and association rules.

Supervised Learning

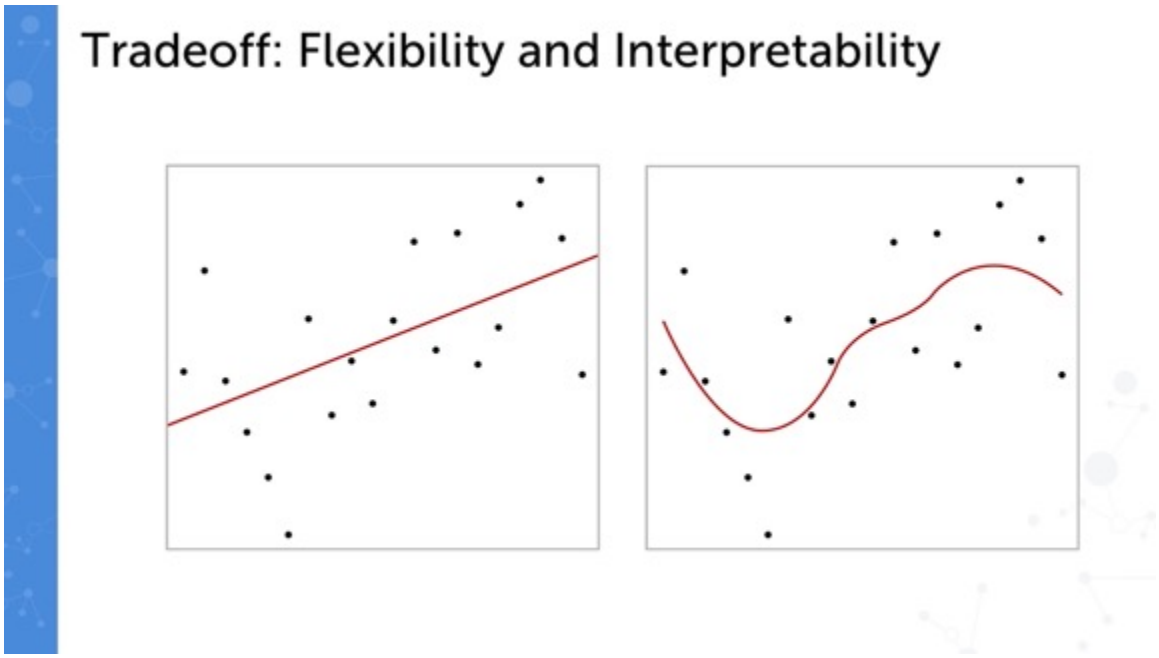
	Response variable is...
Regression	quantitative
Classification	categorical, qualitative

Supervised learning can be further broken down into regression methods and classification methods. Regression methods are used when the response variable is quantitative. So linear regression is an example of a regression method. Classification methods are used when the response variable is categorical, or qualitative. So we want to classify or categorize each data point. Despite its name, logistic regression is an example of a classification method.



However, the boundary between these two approaches can be fuzzy. Many classifications methods-- including logistic regression-- can be used to estimate a probability that a particular data point falls into a particular category. Those probabilities can be thought of as a quantitative response variable. Or you might have a quantitative response variable, for example, sales volume. But maybe you're mainly interested in whether it's high or low. In that case, you might turn your regression problem into a classification problem.

And finally, many methods of supervised learning-- including k nearest neighbors-- apply to both regression and classification problems.

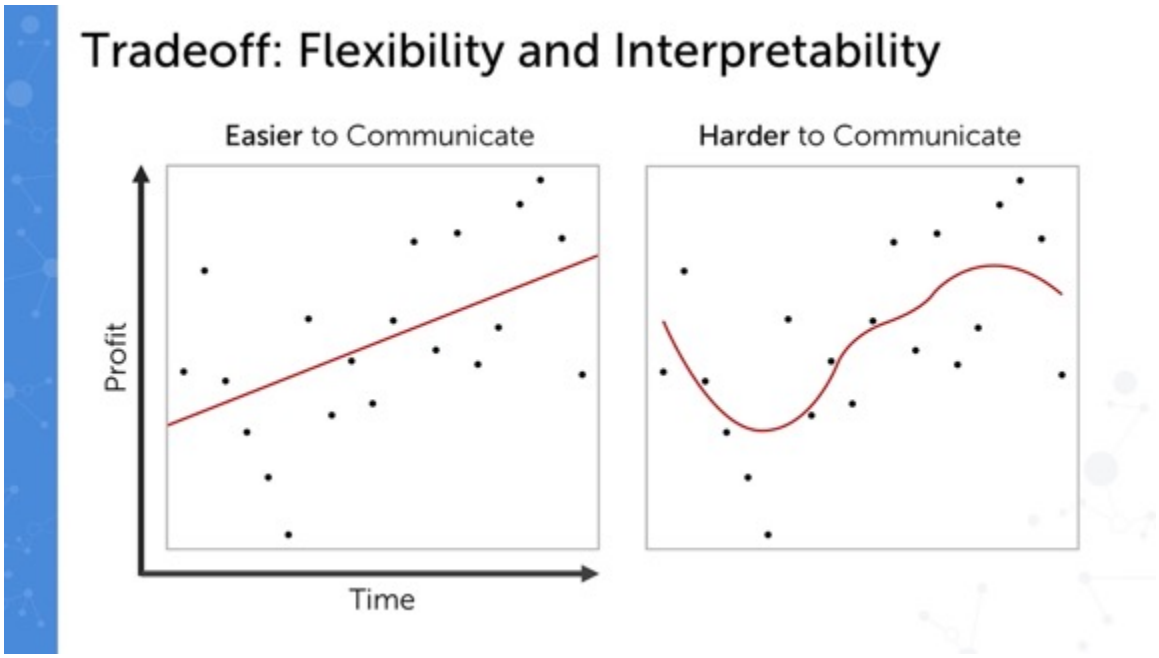


After you know which general category of method you want to use, you'll want to consider the trade off between flexibility and interpretability. More flexible methods have a bigger variety of model shapes that they can produce, which means they can fit the data better. For example, here we have a single data being fit by two different methods. On the left, it's being fit by linear regression, which is only fitting a model shape that looks like a line. On the right, the same dataset is being fit by a lowest smoothing spline, which can fit a variety of curved shapes to the data. That means it can curve around and fit the data better.

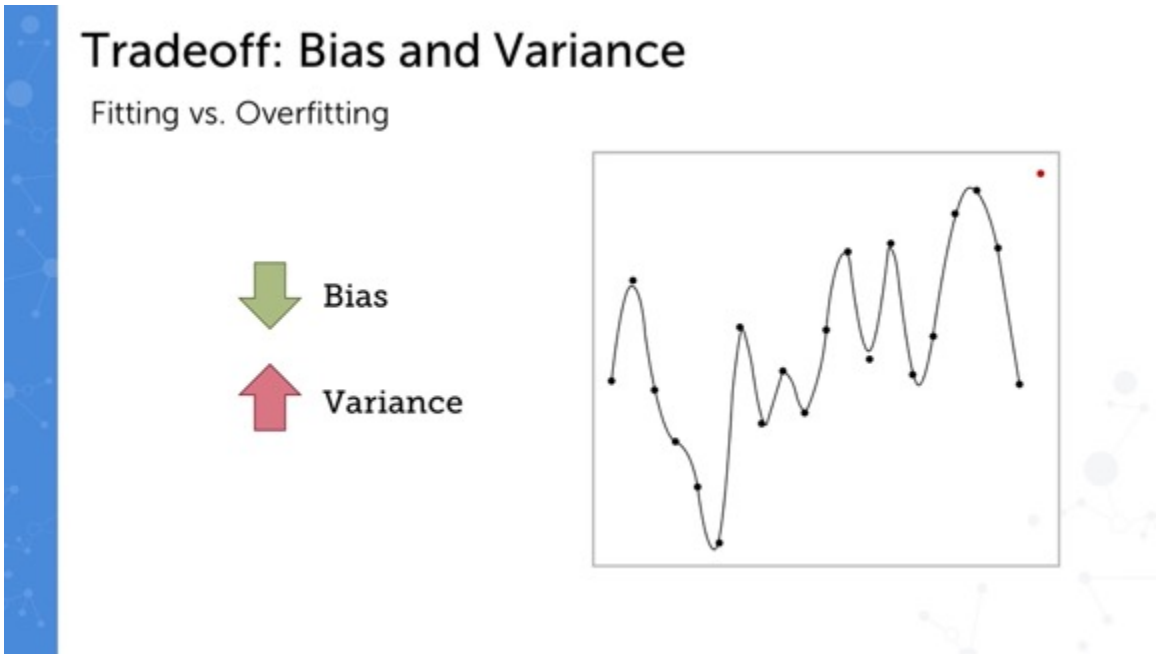
NOTES:

R Code:

```
fit = loess(y ~ x)
xvals = seq(1, 20, .05)
predvals = predict(fit, xvals)
lines(xvals, predvals, lwd=2, col="red")
```



Fitting the data well is good. However, less flexible methods are often easier to interpret. This can be important for communicating your model to others, as well as for understanding the process that underlies your data. For example, suppose that in this dataset, the y-axis represents profit and the x-axis represents the amount of time that a particular store has been open. It might be more useful to know the general trend that as the store has been open longer, the profit tends to increase, as shown by the linear regression. Compared to the picture on the right, knowing that the profit started out decreasing but then increased and then squiggled around some.



Even if your highest priority is the accuracy of the model's predictions, rather than interpreting the model, you might not want the most flexible method. That's because of the trade off between bias and variance. I like to think of this as the balance between fitting the data versus overfitting the data.

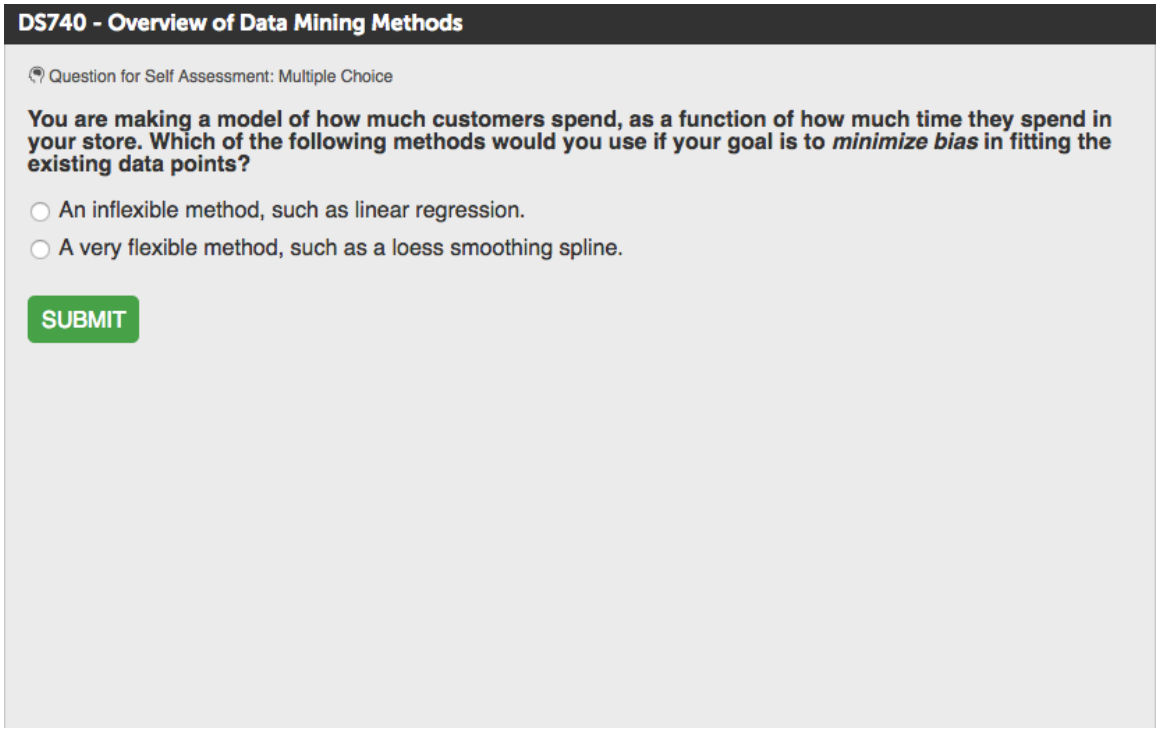
The idea is that if you have a very flexible model, like the interpolating spline shown here, it will fit the data very well, which gives you low bias. However, it's also very sensitive to the specific values in this data set. That means it has high variance, and it means it's likely to be bad for estimating the value of a new data point.

NOTES:

R Code:

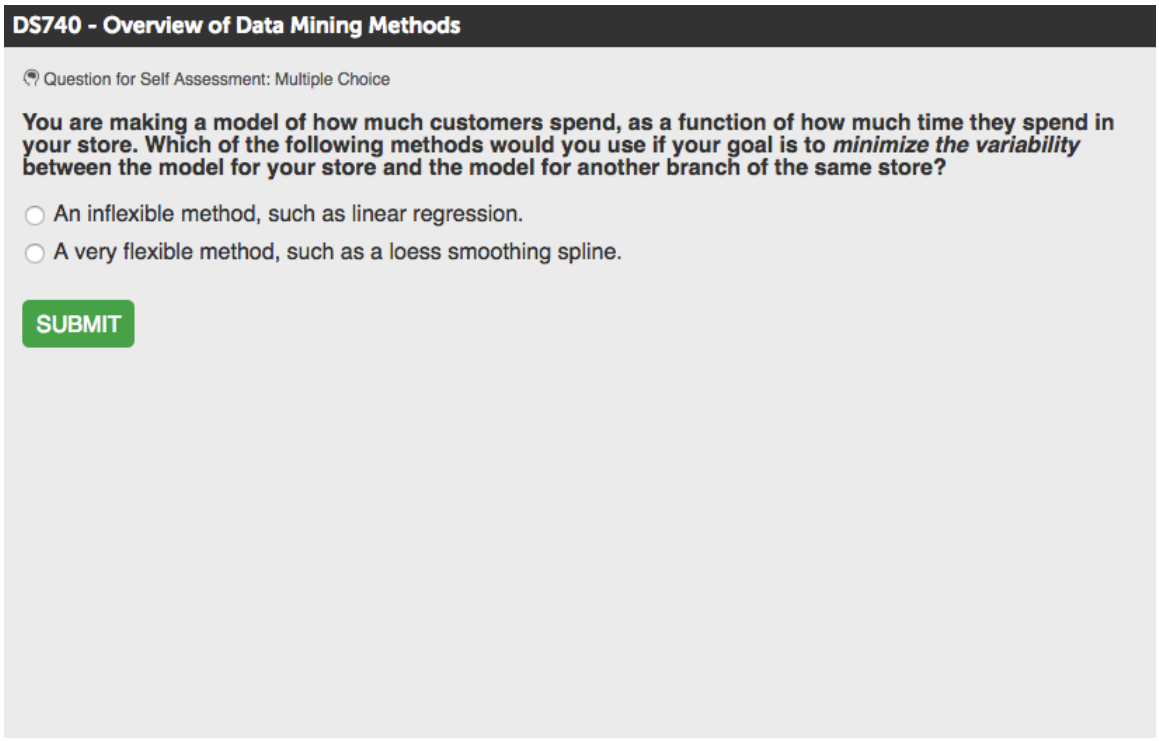
```
library(splines)
fit = interpSpline(y~x)
plot(x, y, pch=19, yaxt="n", xaxt="n", xlim=c(1,20),
ylim=c(49,52.5))
par(new = T)
plot(fit, xlim=c(1,20), ylim=c(49,52.5))
```

Question 1



Answer is at the end of this transcript

Question 2



Answer is at the end of this transcript

Computing Error

Mean ***S***quared ***E***rror: for regression problems

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

Predicted y-value for the ith point

Error rate for classification problems

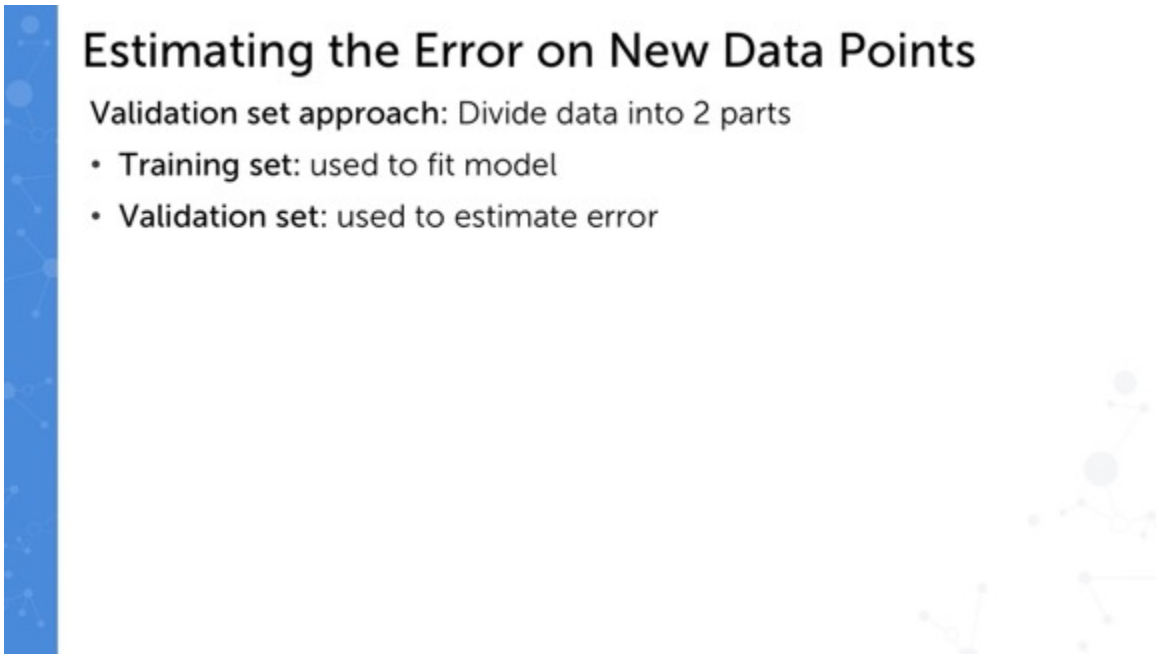
$$\frac{1}{n}\sum_{i=1}^{n} I\left(y_i \neq \hat{y}_i\right)$$

Equals 1 if the ith point is misclassified

Equals 0 if classified correctly

To compute the error for regression-type problems, where the response variable is quantitative, we use the mean squared error. This should look familiar from linear regression. Here, f hat could also be called y sub i hat. It's the predicted y value for the i-th data point.

To compute the error rate for classification problems where the response variable is categorical, we use this formula. Here, i is an indicator variable that equals 1 if y sub i does not equal y sub i hat. That is, if the i-th point is misclassified, and equals 0 if the i-th point is classified correctly.

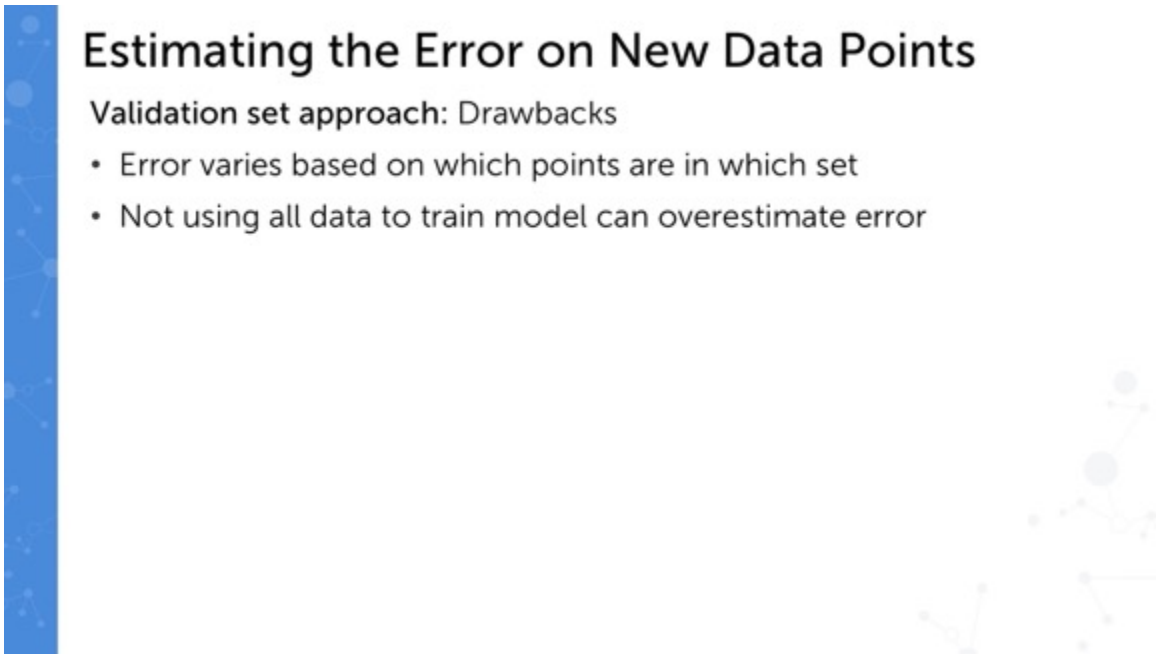


When assessing a model, we want to estimate its error on new data points that weren't used to build the model. One way to do this is with the validation set approach. We divide the data into two parts, a training set, which we use to build the model, and a validation set, which we use to estimate its error. It's important that the training set and validation set be similar in terms of the distribution of the values of the predictor and response variables.

For example, if we're modeling the profits of an ice cream store, we wouldn't want the training set to contain all of the hot days, and the validation set to contain all of the cold days. So we frequently choose a random subset of the data to be in the training set, and the rest of the data to be in the validation set.

NOTES:

See section 5.1.1 in *Introduction to Statistical Learning*.



The validation set approach has two major drawbacks. The first is that the error can vary based on which points are in the training set versus the validation set. The second is that you're not using all of the data to train the model. Perhaps only 1/2 or 2/3 of your data are in the training set. This means you're not using all of the information available, so you can overestimate the error in your model. We'll see how to mitigate both of these drawbacks later on when we discuss cross validation.

Interpreting the model

- Black box model: Cannot be understood by humans
  - Too complex
  - Proprietary
- Creates problems with
  - Knowing when to trust results

After you found a model with a low error rate, it's important to interpret the model. You might be familiar with the term black box, which refers to a model that can't be understood by humans. That might be because it's too complex or simply because it's proprietary, but black box models create problems with knowing when to trust the results.

For example, maybe a model predicts that a particular patient has a high probability of surviving surgery. But did it actually account for this patient's rare blood disorder? You don't know because you don't know what variables the model used or how it used them.

NOTES:

Example of bias in a mortgage lending model:
https://www.cnbc.com/2018/11/27/online-lenders-are-charging-minority-borrowers-more-study-concludes.html

Articles about problems with black box models:
https://arxiv.org/pdf/1811.10154.pdf
https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/6

Interpreting the model

- Black box model: Cannot be understood by humans
  - Too complex
  - Proprietary
- Creates problems with
  - Knowing when to trust results
  - Risk of bias
  - Troubleshooting

- Black box models aren't necessarily more accurate!
  - Look for a good interpretable model first, *and interpret it.*

After you found a model with a low error rate, it's important to interpret the model. You might be familiar with the term black box, which refers to a model that can't be understood by humans. That might be because it's too complex or simply because it's proprietary, but black box models create problems with knowing when to trust the results.

For example, maybe a model predicts that a particular patient has a high probability of surviving surgery. But did it actually account for this patient's rare blood disorder? You don't know because you don't know what variables the model used or how it used them.

NOTES:

Example of bias in a mortgage lending model:
https://www.cnbc.com/2018/11/27/online-lenders-are-charging-minority-borrowers-more-study-concludes.html

Articles about problems with black box models:
https://arxiv.org/pdf/1811.10154.pdf
https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/6

Good interpretations provide insight

- Reveal information about variables/underlying process
- Non-data scientists can understand and provide suggestions
- Suggest strategies for influencing the response variable
- Suggest hypotheses for future research

- Interpretation may be useful even if predictions are insufficiently accurate

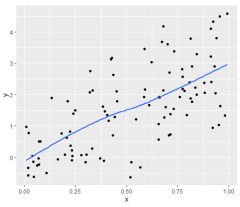
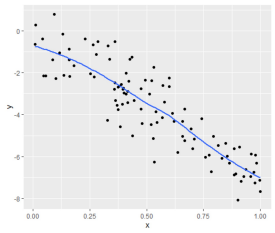
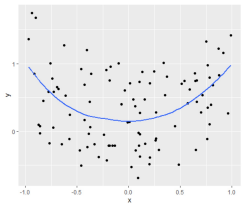
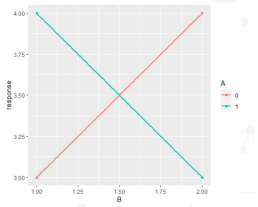
Good interpretations can reveal information about the variables and the underlying process that may have generated the data. They allow non data scientists to understand your model so they can provide suggestions to make it better, like, hey, did you think about including a variable that tells whether the patient has a rare blood clotting disorder?

Good interpretations also can suggest strategies for influencing the response variable. For example, maybe we don't just want to predict the profit of a store, but we want to improve the profit. How might we do that?

And good interpretations can suggest hypotheses for future research or data collection. This means that a good interpretation can be useful even if the predictions are insufficiently accurate for predicting individual observations. They can still tell us something useful about the overall process.

How to write a good interpretation

- What predictor variables are most important?
- Do they have a positive, negative, or curved relationship with the response variable?
- Does the direction of the relationship depend on the value of another variable?

Each predictor variable-- to write a good interpretation, start by looking at which predictor variables are most important in the model and what kind of relationship do they have with the response variable. Is it positive, negative, curved? Or does the direction of the relationship between predictor variable A and the response variable depend on the value of predictor variable—

NOTES:

```
set.seed(101)
x = runif(100)
y = 3*x + rnorm(100)
gf_point(y ~ x) %>%
gf_smooth(y ~ x)
```

```
set.seed(102)
x = runif(100)
y = -7*x + rnorm(100)
gf_point(y ~ x) %>%
gf_smooth(y ~ x)
```

```
set.seed(103)
x = runif(100, -1, 1)
y = x^2 + rnorm(100, 0, .5)
gf_point(y ~ x) %>%
gf_smooth(y ~ x)
```

```
library(ggformula)
mydf = data.frame(A=factor(c(0,0,1,1)), B=c(1,2,1,2), response=c(3,4,4,3))
```

```
gf_point(response ~ B, col =~ A, data = mydf) %>%
gf_smooth(response ~ B, col =~ A, method = "lm", data = mydf)
```

Questions about the relationships

- Do the relationships make sense based on the context of the data?
- Do they agree with previous studies?
- What do the relationships suggest about how to improve the response variable?

Then to take the interpretation deeper you can ask, do the relationships make sense based on the context of the data? Do the directions of the relationships agree with previous studies about similar data sets? And what do the relationships suggest about how to improve the response variable?

You have to take these suggestions with a grain of salt because you don't know if any associations in your data actually represent causal relationships or not, but these suggestions can still provide good hypotheses that you can then use for future investigations.

Good examples

"Higher quality wines tend to have higher alcohol percentages than average quality wines. Additionally, the levels of volatile acidity are 20% higher in average quality wines compared to high quality wines. **This would make sense** given higher volatile acidity levels can contribute to an unpleasant, vinegar taste in the wine."

Here's an example of an interpretation that does a good job of putting its results into a real world context explaining why this makes sense.

Good examples

> “The employees most likely to leave are newer employees who are likely not yet very established within the company. Second, stock options could possibly be **explored as a retention tool**, as it appears that those without stock options are more likely to leave. Third, there is evidence that employees who work overtime are more likely to leave the company.”

Here's an interpretation that does a good job of explaining how these results could be used to affect a change in the response variable, in this case, improving employee retention.

Summary One

Three types of data mining approaches:

- Regression
- Classification
- Unsupervised Learning

The method that best fits the training data is not always the best to use.

- Less flexible methods are often easier to interpret.
- Less flexible methods often have less variance, so they offer a more reliable fit to new data points.

Summary Two

The error of a method can be computed using:

- the MSE: for regression problems.

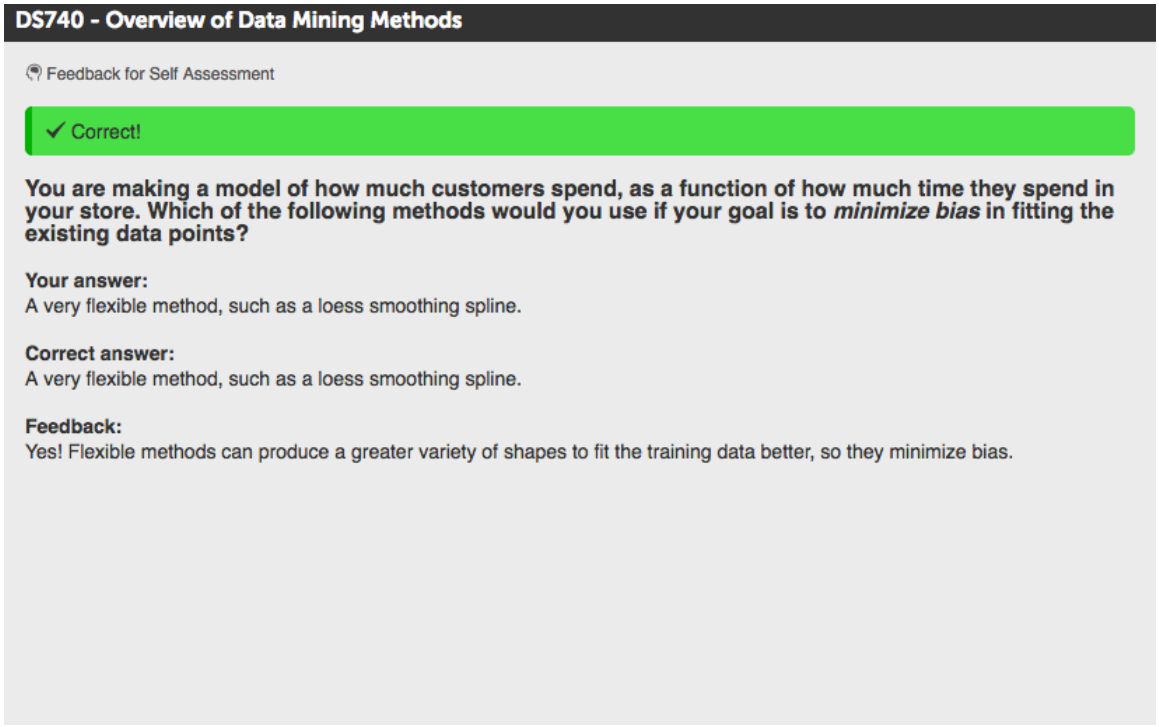
  or

- the proportion of misclassified points: for classification problems.

The validation set approach is one way to estimate a method's error on new data points.

To interpret a model, describe the relationship between the response and the most important predictors, and relate it to the context of the data.

Question 1 Answer



Question 2 Answer