ISLap - Conceptual exercises from chapter 2

Gustavo S. Garone

2025-04-27

Exercise 1

> For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to bebetter or worse than an inflexible method. Justify your answer.

Part a

> the sample size n is extremely large, and the number of predictors p is small.

In this case, we'd expect a flexible model to perform better in this case since overfitting is less of a concern with few predictors.

Part b

> The number of predictors p is extremely large, and the number of observations n is small.

In contrast to (a), with a lot of parameters, we should use an inflexible model as to avoid overfitting.

Part c

> The relationship between the predictors and response is highly non-linear.

With a highly non-linear relationship, we should use more flexible models as to lower the bias.

Part d

> The variance of the error terms, i.e. $\sigma^2 = \text{Var}(\epsilon)$, is extremely high.

In contrast to (c), with a high variance of the error terms, we should aim at using less flexible models as highly flexibles one might overfit to the data and it's errors.

Exercise 2

> Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

Part a

> We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

This is an inference and regression problem, as we are interested in understanding factors (inference) that affect a numerical value (CEO's salaries). We are using data from $n = 500$ companies to learn how the $p = 3$ parameters (profit, number of employees, and industry) affect CEO salary.

Part b

> We are considering launching a new product and wish to know whether it will be a success or afailure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

This is a classification problem concerned with predicting whether the company will fail or not. We are using data from $n = 20$ similar products and for each of them we have $p = 13$ parameters that might affect the company's outcome.

Part c

> We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

This is a regression problem concerned with predicting a numerical (percentual) value. We have data from the $n = 52$ weeks of 2012 and $p = 3$ parameters.

Exercise 3

> We now revisit the bias-variance decomposition.

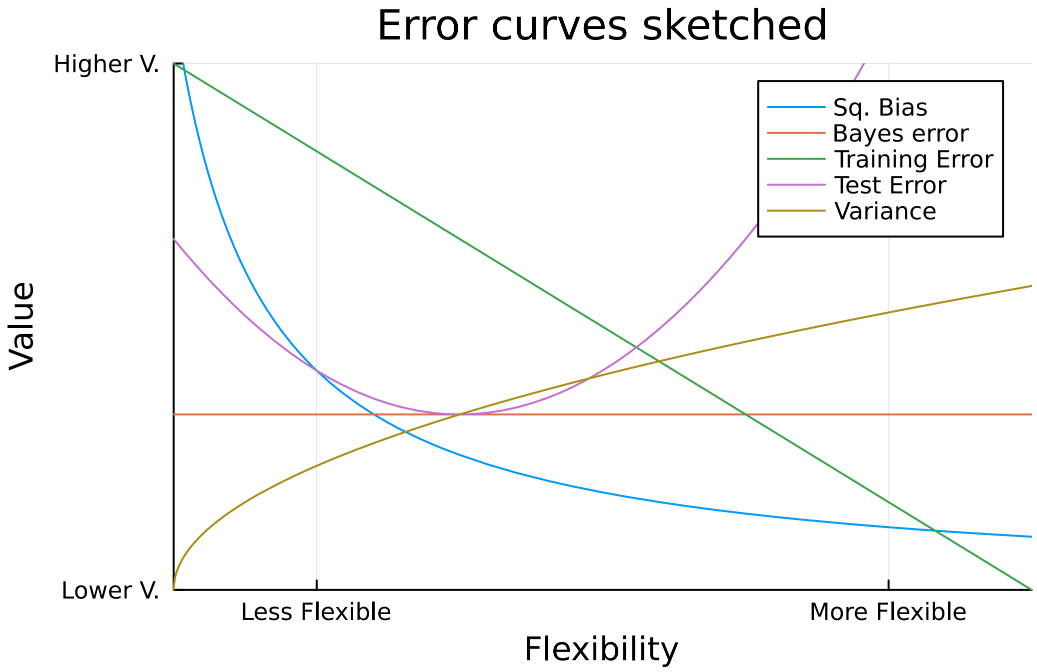
Part a

Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

```julia
using Plots

# Dummy curves for representation, ignore actual form
berror(x) = 1 # Fixed error rate
sbias(x) = 1/(x+0.3) # Decreasing curve
trainerror(x) = 3 - x # Monotonically decreasing
testerror(x) = (x-1)^2 + 1 # U shaped, minimum at bayes error.
variance(x) = sqrt(x) # Incrasing curve


plot(sbias, label="Sq. Bias", ylabel="Value", xlabel="Flexibility",
     title="Error curves sketched", xlims=(0,3), ylims=(0,3),
     xticks=([0.5,2.5], ["Less Flexible", "More Flexible"]),
     yticks=([0,3], ["Lower V.", "Higher V."]))
plot!(berror, label="Bayes error")
plot!(trainerror, label="Training Error")
plot!(testerror, label="Test Error")
plot!(variance, label="Variance")
```



Part b

> Explain why each of the five curves has the shape displayed in part (a).

The squared bias is expected to decrease with more flexible models since flexibility allows for the model to better fit the data. On the other hand, such fitting exacerbates the model's variance (bias-variance tradeoff). As model flexibility increases, the training decreases accordingly since more flexible models better fit the data. In that regard, the test error is expected to decrease to a certain point (never below the fixed Bayes error) before it starts to rise again due to overfitting.

Exercise 4

> You will now think of some real-life applications for statistical learning.

Part a

> Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

1. Classifying skin abnormalities as cancerous or not. In a purely medical scenario, this is a prediction problem. In more academic settings, where we are concerned with what factors predict skin cancer, this is an inference problem. The response is "true" or "false" for labeling an anbnormality as cancerous os not. The predictors in a more abstract sense are color variances, shapes, sizes and patient data. In a more technical sense, the image predictors are the image data: pixel position and (usually RGB) values.

2. Describing the success of an upcoming movie. We might break the classification in five categories, such as "Blockbuster", "Successfull", "Middling", "A flop" and "Complete failure". These categories might be subjective, based on input of other movies' classification provided by the trainer, or more objective, based on factors as financial return or critical acclaim. This could also be approached as either a classification or prediction problem, depending on your goal. Some possible predictors include: total budget, budget percentage per area (filming, scripting, marketing, etc.), star actor count, director experience (quantified in movies made), studio's average movie score on platforms such as IMDB, premiere date.

3. Classifying college athletes as potential professional athletes. This is a binary classification problem ("Yes/no"), usually in a prediction setting. Some usefull predictors might be performance in the last season based on sport-specific metrics (scores, game presence, consistency), weekly training load, personal information (height, weight), years of experience playing the sport competitively. It's a good idea to train separate models for each sport, e.g., height matters more in basketball than other sports.

Part b

> Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

1. Predicting an applying student's final GPA when they finish college. This would be a prediction problem for most colleges, but could be an inference problem in social studies. The measured response is their final GPA score, while some predictors might be high school grades, wether they attended a private or public high school, their household income and their performance on the SAT or other standardized test.
2. Predicting the budget allocation of a household based on the month of the year. This is a prediction problem. We expect as response a vector containing the percentual allocation of total budget for each area, such as groceries, transportation, bills, savings, etc. The predictors might be previous budget allocations, described as a time series across multiple years, as well as inflation (or official inflation predictors).
3. Infering the chance of victory of a game based on the gamestate. This might be a prediction problem for a (e)sports setting, or an inference problem if we are trying to optimise game strategies. The response is a value betweem 0 and 1 (probability of victory), with predictors like score, player positions and other game-specific measurements. Famous examples of machine learning applied to games include the infamous IBM-Kasparov chess match and sophisticated AI systems in video-games.

Part c

> Describe three real-life applications in which cluster analysis might be useful.

Cluster analysis is usefull for classification problems. Some applications include finding optimal ways to create city districts based on planner's criteria, issue voting prediction for targeted campaigning and predicting disease variant for illnesses like dengue fever for targeted vaccination.

Exercise 5

> What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under whata circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

More flexible approaches make better use of a large set of data and allow for more accurately capturing complex patterns, since small data sets would easily result in overfitting. Thus, the use of more flexible models, such as neural networks, allows for reduced bias. That, however, comes with the tradeoff that new data points will likely cause a larger error when comparing to the very precisely data-fit model, which results in a large variance (the bias-variance tradeoff, overfitting). Less flexible models, on the other hand, might have higher bias, specially with very non-linear phenomena, but are more resilient to variance and easier to interpret, which might be important in research applications.

Exercise 6

> Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a non-parametric approach)? What are its disadvantages?

Parametric models depend on parameters for finding the best estimator for the desired result. In statistics, this is called parametric inference (where we are concerned with estimating a parameter (or vector of parameters), θ through inference). Parametric approaches typically require less data to work properly, but might never be able to match the true form of f due to the assumptions made for the form of the function we are estimating, althought this makes such models usually more interpretable. On the other hand, non-parametric models are free to match any given form of f but risk considerable overfitting when not given enough data, which might be a serious limitation depending on the use case.

Exercise 7

> The table below provides a training data set containing six observa-tions, three predictors, and one qualitative response variable.

Obs	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Part a

> Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

Obs	$d(\cdot, (0,0,0))$
1	3
2	2
3	$\sqrt{10}$
4	$\sqrt{5}$
5	$\sqrt{2}$
6	$\sqrt{3}$

Part b

What is our prediction with $K = 1$? Why?

With $K = 1$, our prediction would be green, since the closest point to $(0, 0, 0)$ is observation 5, green.

Part c

What is our prediction with $K = 3$? Why?

With $K = 1$, our prediction would be red, since the 3 closest points to $(0, 0, 0)$ are observations 5 (green), 2 (red) and 6 (red).