Introduction to prediction

Gianluca Campanella

Contents

Prediction and loss functions

Guessing values

- Y = 'time it takes you to get to work in the morning'
- You have some realisations $y_1, y_2, \ldots$ collected over time
- You want to predict the value of Y tomorrow

Guessing values

- Y = 'time it takes you to get to work in the morning'
- You have some realisations $y_1, y_2, \ldots$ collected over time
- You want to predict the value of Y tomorrow

How do you do this?

If you prefer, what's the optimal point forecast for Y?

Loss functions

Before you can answer, you need a loss function that…

- Measures how big an error you're making with your guess g
- Can be minimised to obtain the 'best' g

Loss functions

Before you can answer, you need a loss function that…

- Measures how big an error you're making with your guess g
- Can be minimised to obtain the 'best' g

Mean squared error	$\text{MSE}(g) = \mathbb{E}\left[(Y - g)^2\right]$
Mean absolute error	$\text{MAE}(g) = \mathbb{E}[\lvert Y - g \rvert]$

Towards prediction…

Usually we have at least another variable X that we believe to be related to Y…

Towards prediction...

Usually we have at least another variable X that we believe to be related to Y...

Idea

Using some function f of X, we should be able to predict Y ‘better’ (i.e. reduce the mean error) than by ignoring it

$$g \;\to\; f(X) \quad \text{and thus} \quad \mathrm{MSE}(f) = \mathbb{E}\left[(Y - f(X))^2\right]$$

What should f be?

Consider the decomposition

$$Y|X = f^\star(X) + \varepsilon$$

- $f^\star$ is the optimal prediction (conditional on knowing X)
- ε is a random variable (since $f^\star$ is not)
- $\mathbb{E}[\varepsilon] = 0$ without loss of generality

What should f be?

For the MSE, it can be shown that

$$f^{\star}(x) = \mathbb{E}[Y \mid X = x]$$

$f^{\star}$ is what we'd like to know when we want to predict Y given X

…but can we?

Bias-variance trade-off

Bias-variance trade-off

Suppose that…

- The 'true' regression function is $f^\star$
- We have to make do with some suboptimal f

Let's start by expanding…

$$
\begin{aligned}
(Y - f)^2 &= (Y - f^\star + f^\star - f)^2 \\
&= [(Y - f^\star) + (f^\star - f)]^2 \\
&= (Y - f^\star)^2 + 2(Y - f^\star)(f^\star - f) + (f^\star - f)^2
\end{aligned}
$$

Bias-variance trade-off

Now take the expectation...

$$\mathbb{E}\left[(Y - f^\star)^2 + 2(Y - f^\star)(f^\star - f) + (f^\star - f)^2\right]$$

Since $Y - f^\star = \varepsilon$ and $\mathbb{E}[\varepsilon] = 0$...

- $\mathbb{E}\left[(Y - f^\star)^2\right] = \mathbb{V}[\varepsilon]$
- $\mathbb{E}[Y - f^\star] = \mathbb{E}[\varepsilon] = 0$
- $\mathbb{E}\left[(f^\star - f)^2\right] = (f^\star - f)^2$ (non-random)

Bias-variance trade-off

$$\mathrm{MSE}(f) = \mathbb{V}[\varepsilon] + (f^\star - f)^2$$

Variance $\mathbb{V}[\varepsilon]$

- Doesn't depend on f, just on 'how hard' it is to predict $Y \mid X = x$
- It's the unpredictable, irreducible fluctuation around even the best prediction (randomness rules our lives!)

Bias-variance trade-off

$$\mathrm{MSE}(f) = \mathbb{V}[\varepsilon] + (f^\star - f)^2$$

Bias $(f^\star - f)^2$

- It's the 'extra error' we get from not knowing $f^\star$
- It's also the amount by which we are systematically off

Bias-variance trade-off

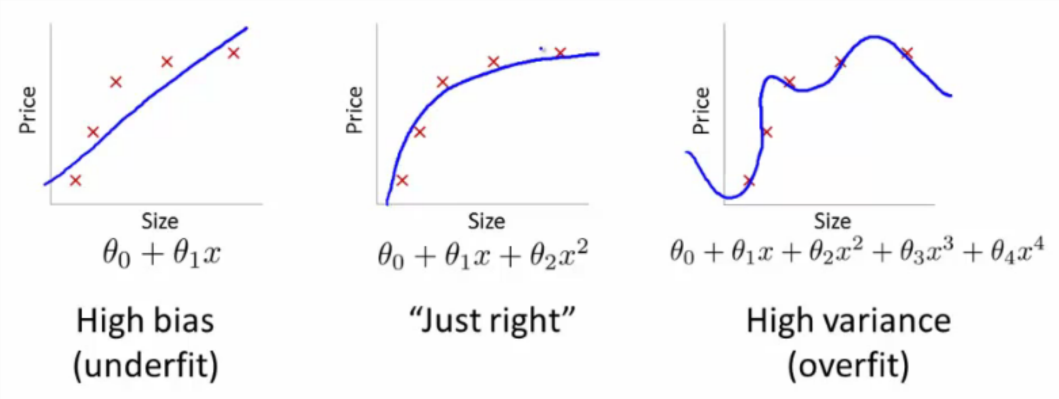
Since f is itself estimated from a sample (it's actually $\hat{f}$), we have…

- The irreducible variance due to the stochastic process
- The bias in approximating $f^\star$ using f
- The additional estimation variance of $\hat{f}$

Consistent methods

- Bias and estimation variance $\rightarrow 0$ as the sample size increases
- Different consistent methods may converge at different rates

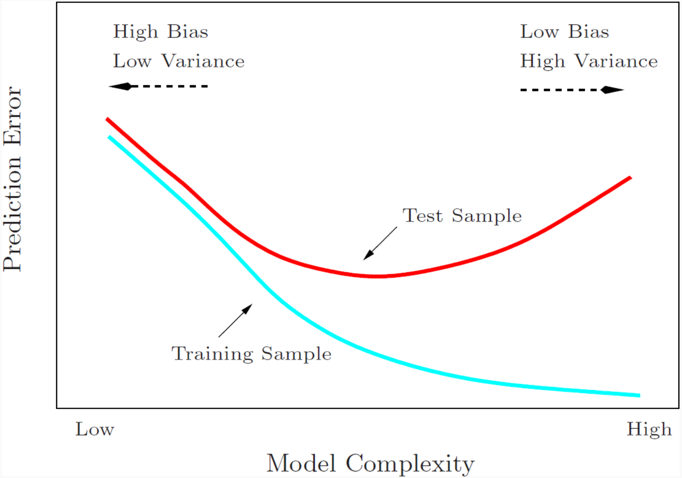
Bias-variance trade-off



From Andrew Ng's *Machine Learning* course

Generalisability

Bias-variance trade-off and generalisability



From *The Elements of Statistical Learning*

Cross-validation

General idea

- Fit several models on subsets of the data
- Measure performance of each
- Compute the mean performance

k-fold cross-validation

- Split the data into *k* groups (a.k.a. ‘folds’)
- Repeat for each fold:
    - Fit the model using all but the selected fold
    - Measure performance on the selected fold
- Compute the mean performance across folds

Regularisation

- Penalise 'large' coefficients by shrinking them
- Helps avoid overfitting
- Requires tuning of an additional parameter α representing the 'weight' of the penalty (relative to the prediction error)

L_1	LASSO	$\sum_j \lvert\beta_j\rvert$
L_2	Tikhonov or ridge	$\sum_j \beta_j^2$