Linear Regression: Test Error

Probability and Statistics for Data Science

Carlos Fernandez-Granda

NYU COURANT INSTITUTE OF MATHEMATICAL SCIENCES

NYU DATA SCIENCE

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Regression

Goal: Estimate response from features

For example, temperature in Versailles (Kentucky) from temperatures at 133 other locations

Linear regression

Linear minimum MSE estimator of response $\tilde{y}$ given features $\tilde{x}$

$$\ell_{\text{MMSE}}(\tilde{x}) = \Sigma_{\tilde{x}\tilde{y}}^{T} \Sigma_{\tilde{x}}^{-1} (\tilde{x} - \mu_{\tilde{x}}) + \mu_{\tilde{y}}$$

Key question: How accurate is the estimate?

Linear response with additive noise

$$\tilde{y} := \tilde{x}^T \beta_{\text{true}} + \tilde{z}$$

Noise $\tilde{z}$ with variance σ^2 independent from the features

For simplicity, everything is centered to have zero mean

What *should* the mean square error be? σ^2

Linear MMSE estimator

$$\tilde{y} := \tilde{x}^T \beta_{\text{true}} + \tilde{z}$$

$$\begin{aligned}
\beta_{\text{MMSE}} &= \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x}\tilde{y}} \\
&= \beta_{\text{true}}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\left[\left(\tilde{y} - \tilde{x}^T \beta_{\text{MMSE}}\right)^2\right] &= \mathrm{E}\left[\left(\tilde{x}^T \beta_{\text{true}} + \tilde{z} - \tilde{x}^T \beta_{\text{true}}\right)^2\right] \\
&= \mathrm{E}\left[\tilde{z}^2\right] \\
&= \sigma^2
\end{aligned}$$

End of story?

No! In practice, we compute linear models from data

Linear regression

Linear minimum MSE estimator of response $\tilde{y}$ given features $\tilde{x}$

$$\ell_{\text{MMSE}}(\tilde{x}) = \Sigma_{\tilde{x}\tilde{y}}^{T} \Sigma_{\tilde{x}}^{-1} (\tilde{x} - \mu_{\tilde{x}}) + \mu_{\tilde{y}}$$

Ordinary-least-squares estimator from dataset
(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)

$$\ell_{\text{OLS}}(x_i) = \Sigma_{XY}^{T} \Sigma_{X}^{-1} (x_i - m(X)) + m(Y)$$

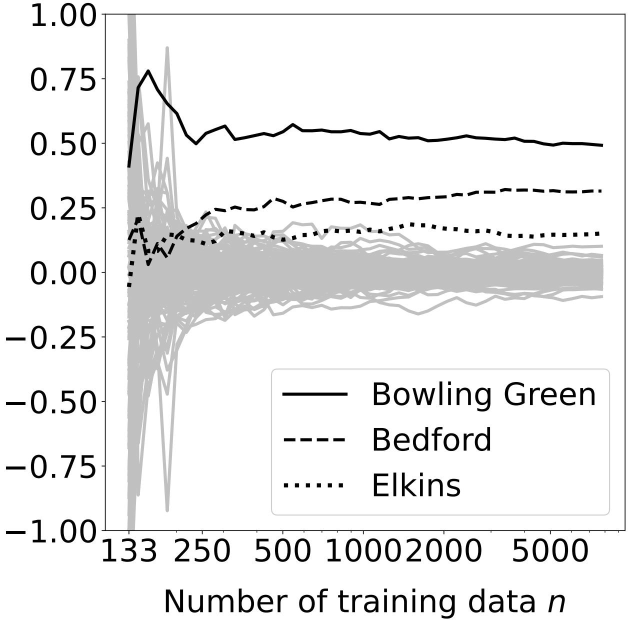
Temperature prediction

Response: Temperature in Versailles (Kentucky)

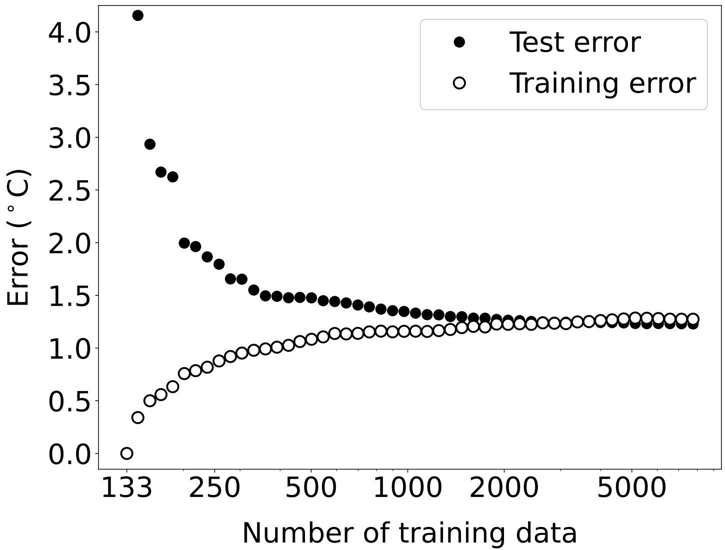
Features: Temperatures at 133 other locations

OLS coefficients (large n)

OLS coefficients



Training and test error



Linear response with additive noise

$$\tilde{y}_{\text{train}} := X_{\text{train}} \beta_{\text{true}} + \tilde{z}_{\text{train}}$$

$$X_{\text{train}} := \begin{bmatrix} x_1^T \\ x_2^T \\ \cdots \\ x_n^T \end{bmatrix}$$

Noise $\tilde{z}_{\text{train}}$ is i.i.d. with variance σ^2 and independent from the features

For simplicity, everything is centered to have zero mean

Test error

Coefficients are estimated from training data:

$$\tilde{y}_{\text{train}} = X_{\text{train}} \beta_{\text{true}} + \tilde{z}_{\text{train}} \quad \Longrightarrow \quad \tilde{\beta}_{\text{OLS}}$$

Test error is computed from test data:

$$(\tilde{x}_{\text{test}}, \tilde{y}_{\text{test}}) \quad \Longrightarrow \quad \tilde{y}_{\text{test}} - \tilde{x}_{\text{test}}^T \tilde{\beta}_{\text{OLS}}$$

$$\tilde{y}_{\text{test}} = \tilde{x}_{\text{test}}^T \beta_{\text{true}} + \tilde{z}_{\text{test}}$$

Noise $\tilde{z}_{\text{test}}$ with variance σ^2 independent from everything else

OLS coefficients

$$\mathrm{E}\left[\tilde{\beta}_{\mathsf{OLS}}\right] = \beta_{\mathsf{true}}$$

$$\Sigma_{\tilde{\beta}_{\mathsf{OLS}}} = \frac{\sigma^2}{n-1}\Sigma_X^{-1}$$

Goal: Understand how coefficient error propagates to test error

Test error

$$\begin{aligned}\tilde{y}_{\text{test}} - \tilde{x}_{\text{test}}^T \tilde{\beta}_{\text{OLS}} &= \tilde{x}_{\text{test}}^T \beta_{\text{true}} + \tilde{z}_{\text{test}} - \tilde{x}_{\text{test}}^T \left(\beta_{\text{true}} + \text{ct}(\tilde{\beta}_{\text{OLS}})\right) \\ &= \tilde{z}_{\text{test}} - \tilde{x}_{\text{test}}^T \, \text{ct}(\tilde{\beta}_{\text{OLS}})\end{aligned}$$

Bad news: OLS coefficients can have **high** variance in certain directions (due to feature collinearity and limited data)

Good news: The features have **low** variance in those directions!

Mean squared test error

$$\tilde{y}_{\text{test}} - \tilde{x}_{\text{test}}^T \tilde{\beta}_{\text{OLS}} = \tilde{z}_{\text{test}} - \tilde{x}_{\text{test}}^T \operatorname{ct}(\tilde{\beta}_{\text{OLS}})$$

$$\begin{aligned} & \mathrm{E}\left[\left(\tilde{y}_{\text{test}} - \tilde{x}_{\text{test}}^T \tilde{\beta}_{\text{OLS}}\right)^2\right] \\ & = \mathrm{Var}\left[\tilde{z}_{\text{test}}\right] + \mathrm{E}\left[\tilde{x}_{\text{test}}^T \operatorname{ct}(\tilde{\beta}_{\text{OLS}}) \operatorname{ct}(\tilde{\beta}_{\text{OLS}})^T \tilde{x}_{\text{test}}\right] \\ & = \sigma^2 + \mathrm{E}\left[\tilde{x}_{\text{test}}^T \operatorname{ct}(\tilde{\beta}_{\text{OLS}}) \operatorname{ct}(\tilde{\beta}_{\text{OLS}})^T \tilde{x}_{\text{test}}\right] \end{aligned}$$

Test error

Assuming $\Sigma_{\tilde{x}_{\text{test}}} = \Sigma_X$

$$
\begin{aligned}
&\mathrm{E}\left[\tilde{x}_{\text{test}}^T \, \text{ct}(\tilde{\beta}_{\text{OLS}}) \, \text{ct}(\tilde{\beta}_{\text{OLS}})^T \tilde{x}_{\text{test}}\right] \\
&= \mathrm{E}\left[\text{Trace}\left(\tilde{x}_{\text{test}}^T \, \text{ct}(\tilde{\beta}_{\text{OLS}}) \, \text{ct}(\tilde{\beta}_{\text{OLS}})^T \tilde{x}_{\text{test}}\right)\right] \\
&= \mathrm{E}\left[\text{Trace}\left(\tilde{x}_{\text{test}} \tilde{x}_{\text{test}}^T \, \text{ct}(\tilde{\beta}_{\text{OLS}}) \, \text{ct}(\tilde{\beta}_{\text{OLS}})^T\right)\right] \\
&= \text{Trace}\left(\mathrm{E}\left[\tilde{x}_{\text{test}} \tilde{x}_{\text{test}}^T \, \text{ct}(\tilde{\beta}_{\text{OLS}}) \, \text{ct}(\tilde{\beta}_{\text{OLS}})^T\right]\right) \\
&= \text{Trace}\left(\mathrm{E}\left[\tilde{x}_{\text{test}} \tilde{x}_{\text{test}}^T\right] \mathrm{E}\left[\text{ct}(\tilde{\beta}_{\text{OLS}}) \, \text{ct}(\tilde{\beta}_{\text{OLS}})^T\right]\right) \\
&= \text{Trace}\left(\Sigma_{\tilde{x}_{\text{test}}} \Sigma_{\tilde{\beta}_{\text{OLS}}}\right) \\
&= \text{Trace}\left(\Sigma_X \frac{\sigma^2}{n-1} \Sigma_X^{-1}\right) \\
&= \frac{\sigma^2}{n-1} \text{Trace}\left(I\right) = \frac{\sigma^2 d}{n-1}
\end{aligned}
$$

Test error

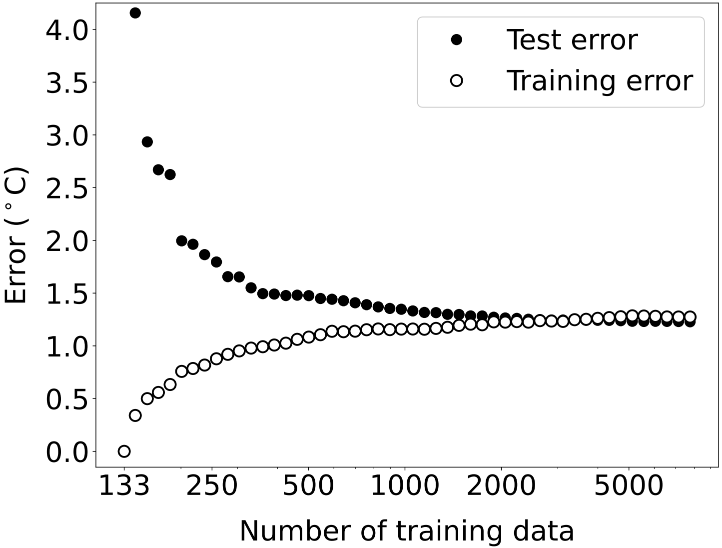
If $\Sigma_{\tilde{x}_{\text{test}}} = \Sigma_X$

$$\mathrm{E}\left[\left(\tilde{y}_{\text{test}} - \tilde{x}_{\text{test}}^T \tilde{\beta}_{\text{OLS}}\right)^2\right] = \sigma^2 + \frac{\sigma^2 d}{n-1}$$
$$= \sigma^2 \left(1 + \frac{d}{n-1}\right)$$

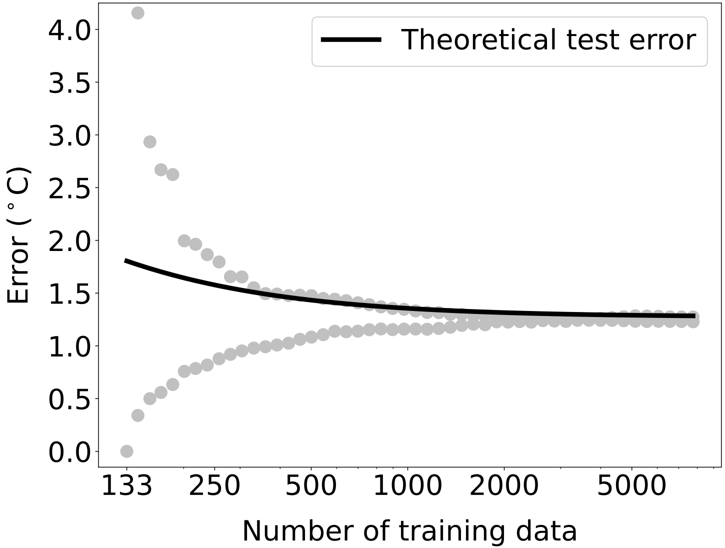
When $n \gg d$? σ^2

When $n \approx d$? $2\sigma^2$? Probably not, because $\Sigma_{\tilde{x}_{\text{test}}} \neq \Sigma_X$!

Temperature prediction



Theoretical analysis



What have we learned?

Test error depends on number of training data n

If $n \gg d$: Generalization to test data

If $n \approx d$: Overfitting
(due to *inaccurate estimation of feature covariance*)