Linear Regression: Training 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 well do we fit the data?

Linear response with additive noise

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

Noise $\tilde{z}$ has variance σ^2 and is independent from the features $\tilde{x}$

For simplicity, everything is centered to have zero mean

What *should* the mean squared 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)

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

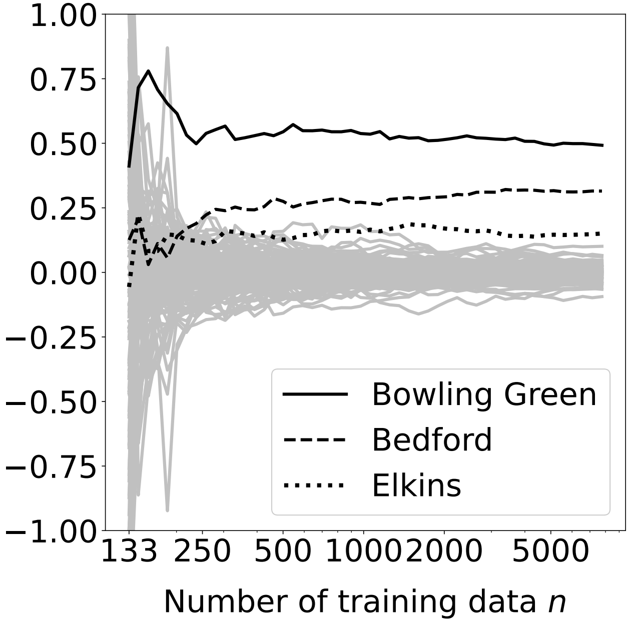
Temperature prediction

Response: Temperature in Versailles (Kentucky)

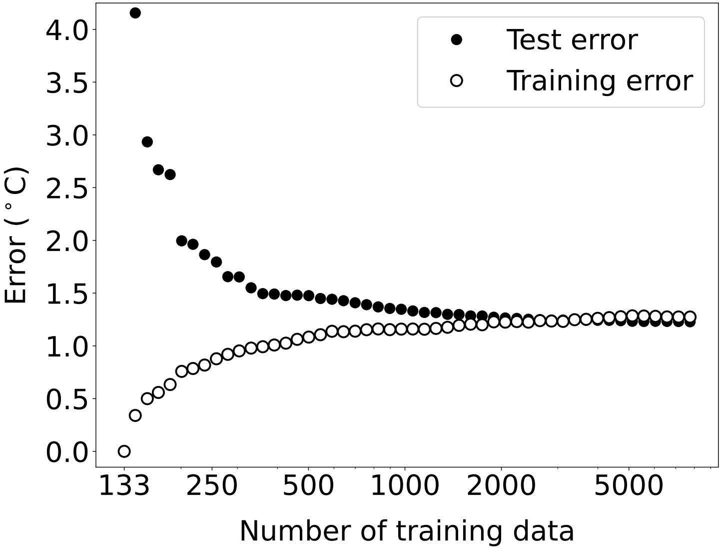
Features: Temperatures at 133 other locations

OLS coefficients (large n)

OLS coefficients



Training and test error



Linear response with additive noise

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

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

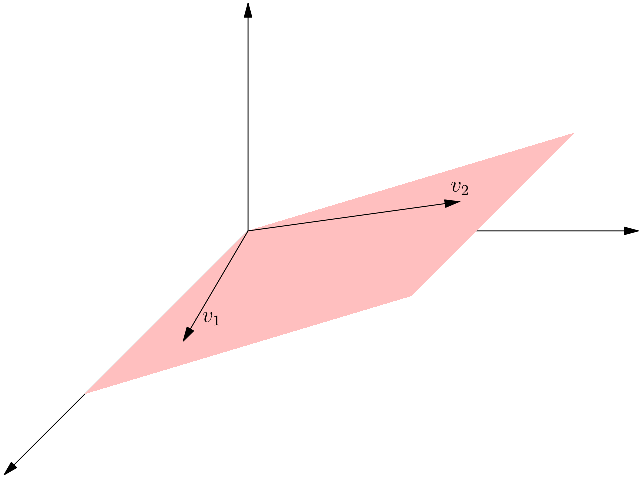
Noise $\tilde{z}_{\text{train}}$ is i.i.d. with variance σ^2

For simplicity, everything is centered to have zero mean

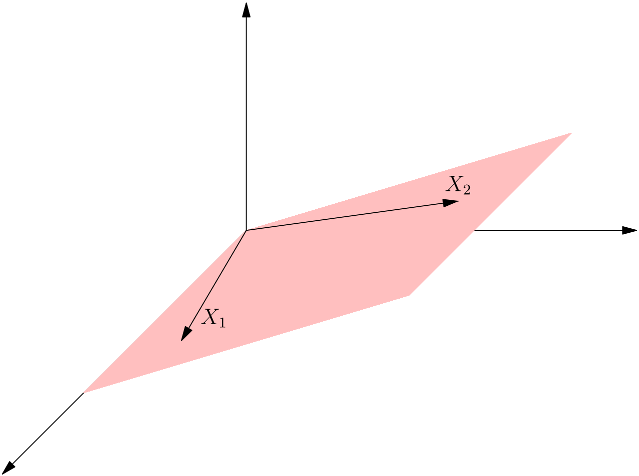
From a linear algebra perspective

$$
\begin{aligned}
X_{\text{train}}\beta &= \begin{bmatrix} x_1[1] & x_1[2] & \cdots & x_1[d] \\ x_2[1] & x_2[2] & \cdots & x_2[d] \\ \cdots & \cdots & \cdots & \cdots \\ x_n[1] & x_n[2] & \cdots & x_n[d] \end{bmatrix} \beta \\
&= \begin{bmatrix} X_1 & X_2 & \cdots & X_d \end{bmatrix} \beta \\
&= \beta[1]\, X_1 + \beta[2]\, X_2 + \cdots + \beta[d]\, X_d
\end{aligned}
$$

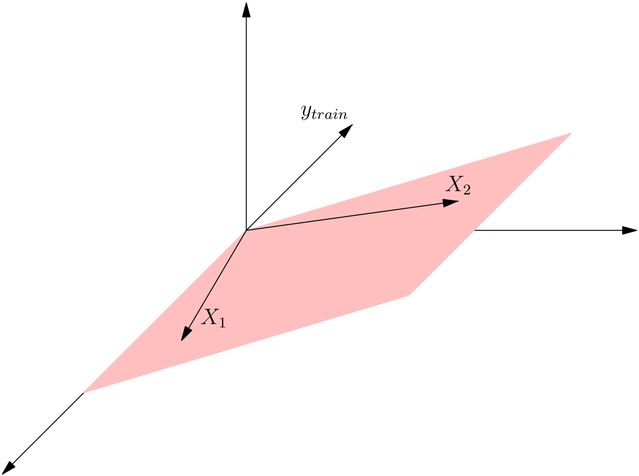
Subspace



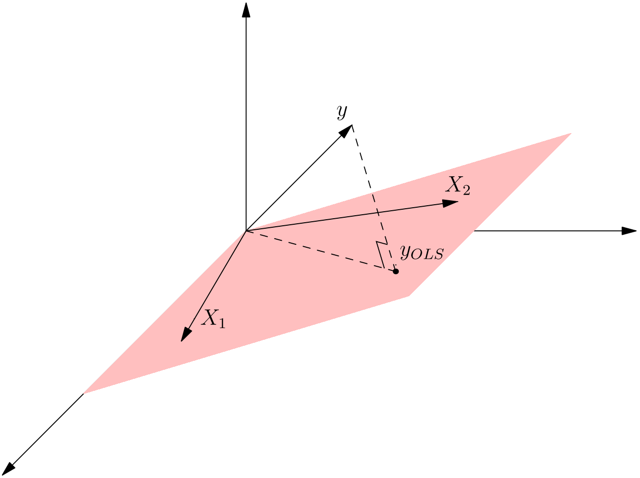
Linear model



OLS estimator?



Projection



Training error

$$\tilde{y}_{\text{OLS}} = \mathcal{P}_{\text{col}(X_{\text{train}})}\, \tilde{y}_{\text{train}}$$

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

$$\begin{aligned}
\tilde{y}_{\text{train}} - \tilde{y}_{\text{OLS}} &= \tilde{y}_{\text{train}} - \mathcal{P}_{\text{col}(X_{\text{train}})}\, \tilde{y}_{\text{train}} \\
&= \mathcal{P}_{\text{col}(X_{\text{train}})^{\perp}}\, \tilde{y}_{\text{train}} \\
&= \mathcal{P}_{\text{col}(X_{\text{train}})^{\perp}} \left(X_{\text{train}} \beta_{\text{true}} \right) + \mathcal{P}_{\text{col}(X_{\text{train}})^{\perp}}\, \tilde{z}_{\text{train}} \\
&= \mathcal{P}_{\text{col}(X_{\text{train}})^{\perp}}\, \tilde{z}_{\text{train}}
\end{aligned}$$

Dimension of $\text{col}(X_{\text{train}})^{\perp}$? $n - d$

Training error? Variance captured by projection $\approx \sigma^2 (n - d)$

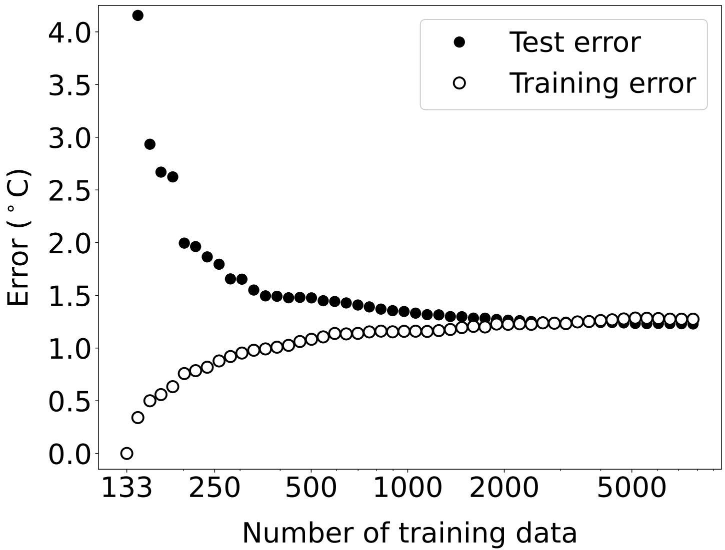
Average training error

$$\text{Average training error} \approx \frac{\sigma^2 (n-d)}{n} = \sigma^2 \left(1 - \frac{d}{n}\right)$$

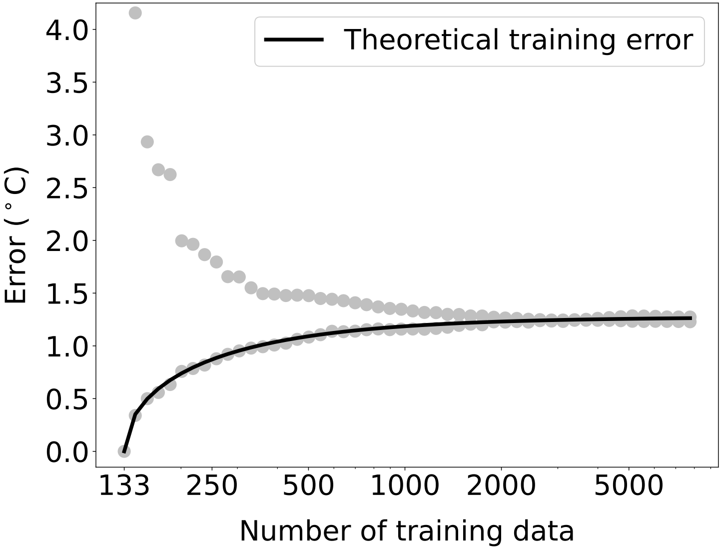
When $n \gg d$? σ^2

When $n \approx d$? Very small!

Temperature prediction



Theoretical analysis



What have we learned?

Training error depends on the number of training data n

If $n \gg d$: no overfitting

If $n \approx d$: overfitting