From Overfitting to Generalization: Regularization in Action

DATA443: Statistical Machine Learning

Surani Matharaarachchi

University of Manitoba

March 2025

Recap: Regression and Classification

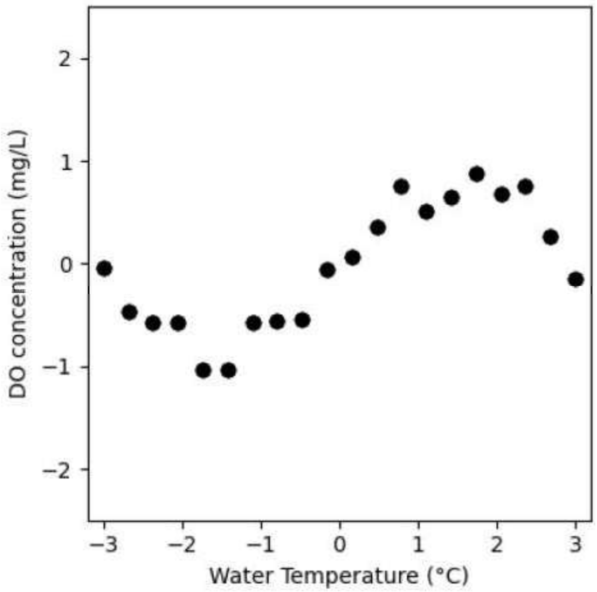
Key Topics Covered:

- Difference between classification and regression tasks.
- Common algorithms: Linear Regression, Logistic Regression, Decision Trees, SVM.
- Evaluating model performance: Accuracy, Precision, Recall, RFmulMSE, R-squared, F1-Score.
- Cross Validation.

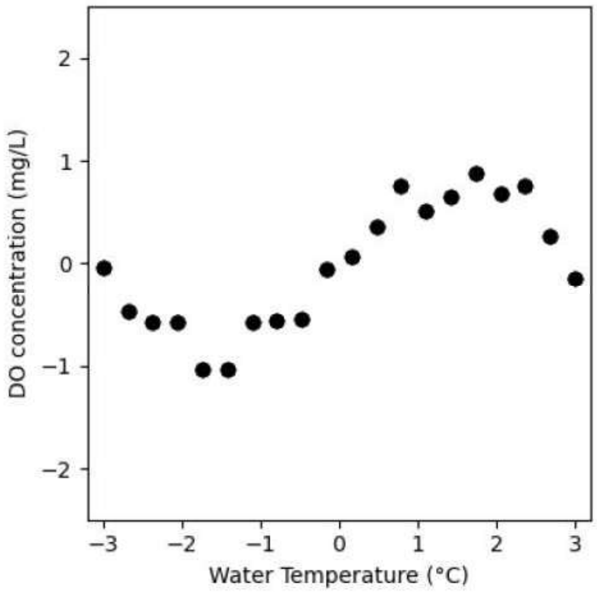
Today's Outline

- Need for Regularization
- What is Regularization?
- Regularization Techniques — L1 and L2
- Hands-On

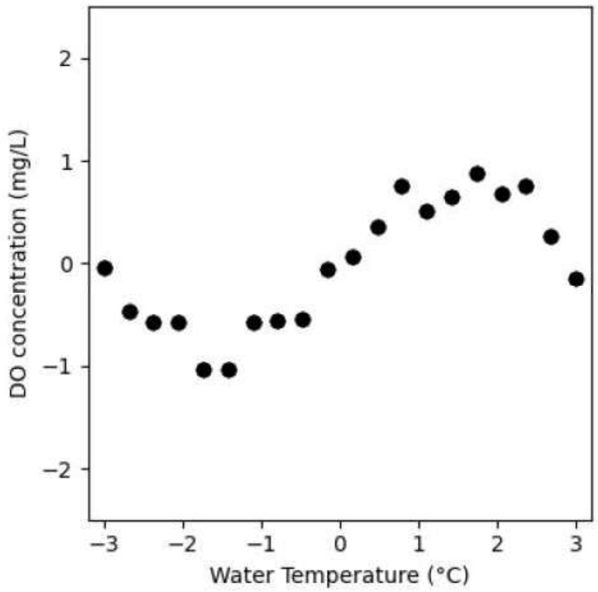
Need for Regularization: Linear Model Example



(a) $\beta_0 + \beta_1 x + \epsilon$



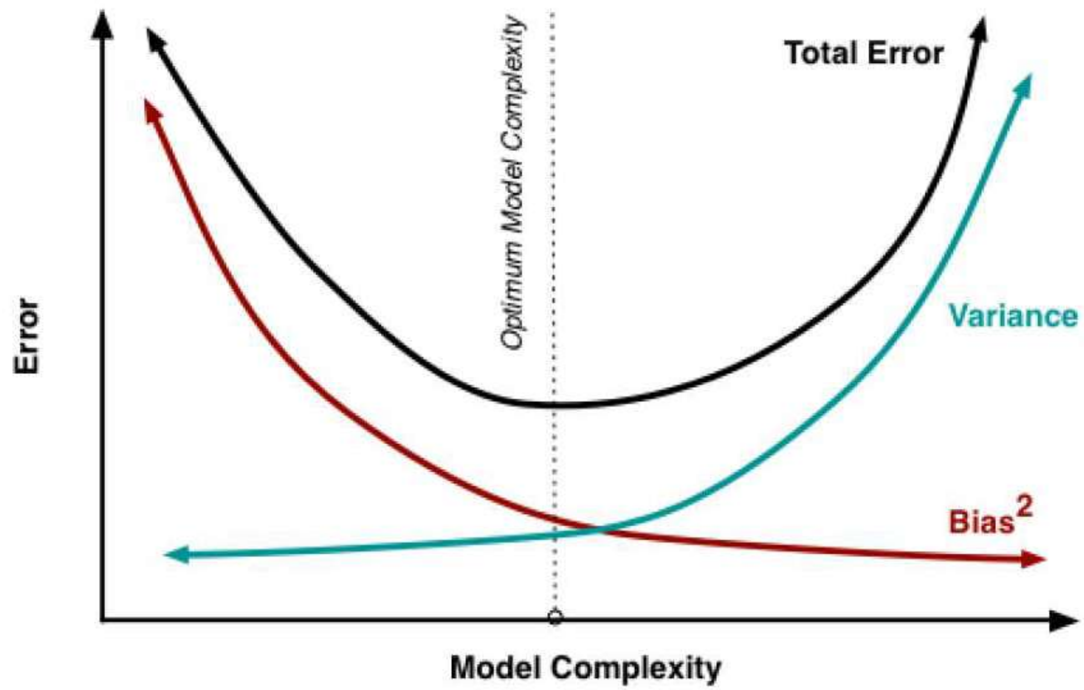
(b) $\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \epsilon$



(c) $\beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_{15} x^{15} + \epsilon$

Bias-Variance Trade-Off

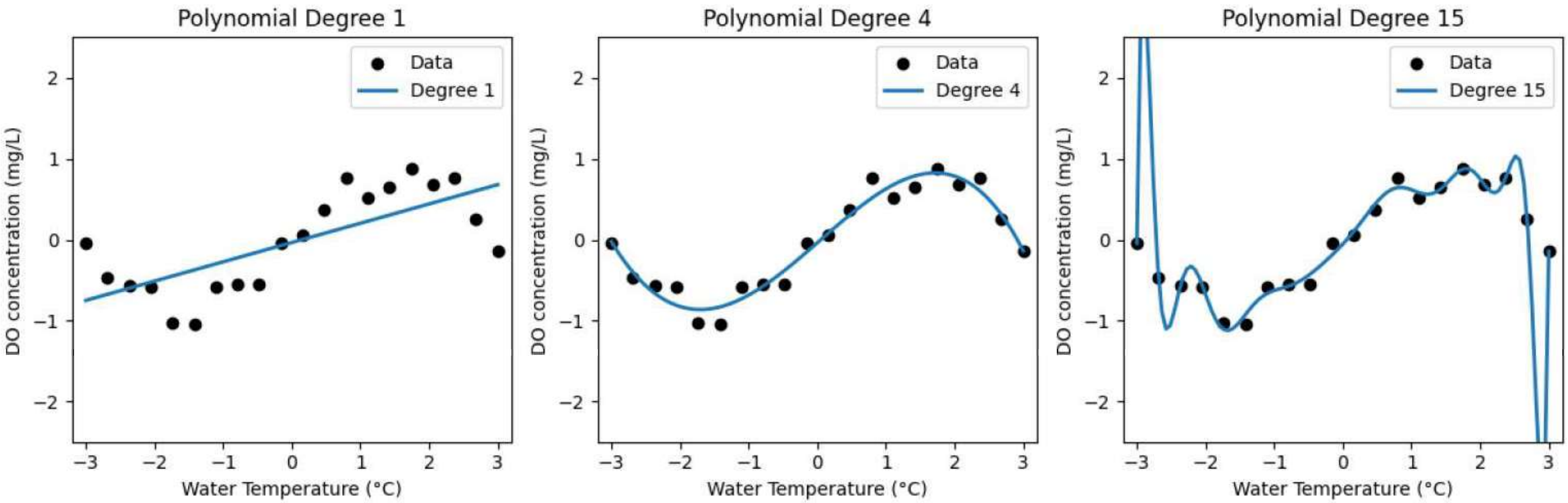
- Bias is the error due to overly simplistic assumptions in the learning algorithm.
- Variance is the error due to the model being too sensitive to small fluctuations in the training data.
- "sweet spot" - a model complex enough to learn patterns but simple enough to generalize



What is Overfitting?

- **Overfitting** is a modeling error that occurs when a machine learning model learns not only the underlying patterns in the training data but also the noise and random fluctuations.
- As a result, the model performs well on the training data but poorly on unseen data.
- Symptoms include high training accuracy but low test accuracy.

What is Overfitting?



Polynomial Degree	**1**	**4**	**15**
Train MSE	0.1735	0.0176	0.0045
Test MSE	0.2021	0.0190	0.6052

Causes of Overfitting

- **Too Complex Model:** Models with a large number of parameters relative to the amount of training data prone to overfitting.
- **Limited Training Data:** A small dataset increases the risk of memorization rather than learning generalizable patterns.
- **Noise in Data:** If the training data contained noise or irrelevant patterns, the model may treat these as if they are genuine features.

Addressing overfitting

Problem:
If we have too many features, the learned model may fit the training set very well, but fail to generalize to new examples.

Solutions:

1. Cross Validation
2. More data
3. Reduce the number of features
   - Manually select which features to keep.
   - Model selection algorithm (later in the course).
4. Regularization
   - **shrinkage** in statistics

Regularization

Regularization involves modifying the loss function L by introducing an additional term that penalizes some specified properties of the model parameters.

$$L_{reg}(\beta) = L(\beta) + \lambda R(\beta), \quad (1)$$

- λ is a scalar that is called **regularization parameter** that gives the weight (or importance) of the regularization term.
- This added penalty term helps to control the complexity of the model and prevent overfitting.

Regularization

- Regularization Methods for Linear Models
    - Ridge Regression (L2 Regularization)
    - LASSO Regression (L1 Regularization)
    - Elastic Net Regularization (Combination of L1 and L2)
- Regularization Methods for Neural Networks (later in the course)
    - L2 Regularization (Weight Decay)
    - L1 Regularization (Sparse Regularization)
    - Dropout
    - Early Stopping
    - Batch Normalization (acts as a form of regularization)
    - Data Augmentation (for increasing generalization)

LASSO Regression ($L1$ Regularization)

To prevent extreme values in the model parameters, we incorporate a regularization term that penalizes large magnitudes. In this case, we use RSS as our loss function.

Regularized Loss Function:

$$L_{LASSO}(\beta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \qquad (2)$$

where $\hat{y}_i = \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}$

Finding the model parameters β_{LASSO} that minimize the L_1 regularized loss function is called **LASSO regression.**

$$\min_{\beta}(L_{LASSO})$$

LASSO Regression: Strengths and Limitations

Strengths

- Prevent overfitting by penalizing large coefficients.
- Shrinks some coefficients to zero and yields sparse models.
- Improves model interpretability.
- Useful in high-dimensional settings where many features are irrelevant.

Limitations

- Performance depends on the regularization parameter.
- Model can be biased.
- Sensitive to outliers.
- Limited in capturing complex, non-linear relationships.
- Can remove useful features.

Ridge Regression ($L2$ Regularization)

Alternatively, we can choose a regularization term that penalizes the squares of the parameter magnitudes.

Regularized Loss Function:

$$L_{Ridge}(\beta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \quad (3)$$

Finding the model parameters β_{Ridge} that minimize the ℓ_2 regularized loss function is called **Ridge regression.**

$$\min_{\beta}(L_{Ridge})$$

Ridge Regression: Strengths and Limitations

Strengths

- Prevent overfitting by penalizing large coefficients.
- Handles multicollinearity.
- Maintains all features.
- More robust to outliers.

Limitations

- Requires careful tuning of the regularization parameter.
- No feature selection.
- Can introduce bias, risking underfitting if λ is too large.
- Can be less interpretable.

Elastic Net Regression

Elastic Net is a regularized linear regression model that combines LASSO ($L1$) and Ridge ($L2$) penalties.

Regularized Loss Function:

$$L_{Elastic\_Net}(\beta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2. \tag{4}$$

To control the balance between L_1 and L_2 penalties:

$$L_{Elastic\_Net}(\beta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda[\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2]. \tag{5}$$

Elastic Net Regression

- Finding the model parameters $\beta_{Elastic\_Net}$ that minimize the regularized loss function is called **Elastic net regression.**

$$\min_{\beta}(L_{Elastic\_Net})$$

Interpretation:

- Balances Ridge and LASSO.
- Useful when features are highly correlated.
- Uses a mixing parameter α to control balance.

Selecting the Optimal λ Value

How to Choose λ?

- $\lambda = 0$ results in no regularization (OLS regression).
- A large λ leads to overly simplified models (high bias, low variance).
- The best λ balances model complexity and generalization.

Common Approaches:

- Cross-validation: Find λ that minimizes validation error.
- Grid search: Test multiple λ values and compare performance.
- Information criteria: Use AIC or BIC to guide selection.

Applying Regularization Techniques to a Water Conservation Dataset

Objective: Predict the target variable **Dissolved Oxygen (DO)** accurately and assess the effect of regularization techniques.

Dataset: Example Dataset for Water Quality Prediction

- Contains 100 observations with 10 scientifically meaningful features.
- Features include temperature, pH, turbidity, nutrient levels, and chemical demand. DO, a key indicator of water quality.

Comparative Analysis: Comparison of Linear, Lasso, and Ridge Regression

Click the link to the code file

Results - Model Performance Summary

Table: Model Performance Summary

	OLS Regression	Ridge Regression	LASSO Regression
Train MSE	2.774	2.923	2.965
Test MSE	3.376	3.097	**2.823**
Train R^2	0.983	0.982	0.982
Test R^2	0.983	0.985	**0.986**

Results - Feature Importance

Table: Feature Importance (Regression Coefficients) for DO Prediction

Feature	OLS Coef	Ridge Coef	LASSO Coef
Temperature (°C)	-2.31	-2.24	-2.14
pH	-0.14	-0.08	**0.00**
Turbidity (NTU)	-2.39	-2.36	-2.23
Conductivity (µS/cm)	-23.52	-2.74	-6.09
Nitrate (mg/L)	-1.65	-1.62	-1.54
Phosphate (mg/L)	-2.39	-1.65	-1.84
BOD (mg/L)	-4.82	-4.50	-2.81
COD (mg/L)	-4.66	-5.44	-3.28
TDS (mg/L)	-8.33	-2.20	**0.00**
Salinity (ppt)	-6.43	-2.63	-0.61

Interpretation of Regression Results

- **Linear Regression:**
  - High accuracy ($R^2 = 0.983$), but sensitive to multicollinearity.
  - Coefficient magnitudes are unstable due to correlated predictors.
- **Ridge Regression:**
  - Slightly lower training accuracy but better generalization (Test $R^2 = 0.985$).
  - L2 penalty stabilizes coefficients by shrinking them-ideal for multicollinearity.
- **LASSO Regression:**
  - Achieved the best test performance ($MSE = 2.82$, $R^2 = 0.986$).
  - L1 regularization induces sparsity by eliminating less important features (e.g., pH coefficient = 0).

Conclusion

- Regularization helps prevent overfitting.
- Regularization improves generalization under multicollinearity.
- **LASSO** is best for interpretability and variable selection.
- **Ridge** is robust when all features are important but correlated.
- **Elastic Net** balances both techniques.

Next Class

Topics:

- Feature Selection Techniques

Homework Assignment

Tasks:

- Implement and compare different regularization techniques on the given real-world dataset.
- More details about the assignment are provided in the course website.

Submission Deadline: Next class session