Cheat Sheet: Linear and Logistic Regression

Comparing different regression types

Model Name	Description	Code Syntax
Simple linear regression	**Purpose:** To predict a dependent variable based on one independent variable. **Pros:** Easy to implement, interpret, and efficient for small datasets. **Cons:** Not suitable for complex relationships; prone to underfitting. **Modeling equation:** $y = b_0 + b_1x$	```python\nfrom sklearn.linear_model import LinearRegression\nmodel = LinearRegression()\nmodel.fit(X, y)\n```
Polynomial regression	**Purpose:** To capture nonlinear relationships between variables. **Pros:** Better at fitting nonlinear data compared to linear regression. **Cons:** Prone to overfitting with high-degree polynomials. **Modeling equation:** $y = b_0 + b_1x + b_2x^2 + ...$	```python\nfrom sklearn.preprocessing import PolynomialFeatures\nfrom sklearn.linear_model import LinearRegression\npoly = PolynomialFeatures(degree=2)\nX_poly = poly.fit_transform(X)\nmodel = LinearRegression().fit(X_poly, y)\n```
Multiple linear regression	**Purpose:** To predict a dependent variable based on multiple independent variables. **Pros:** Accounts for multiple factors influencing the outcome. **Cons:** Assumes a linear relationship between predictors and target. **Modeling equation:** $y = b_0 + b_1x_1 + b_2x_2 + ...$	```python\nfrom sklearn.linear_model import LinearRegression\nmodel = LinearRegression()\nmodel.fit(X, y)\n```
Logistic regression	**Purpose:** To predict probabilities of categorical outcomes. **Pros:** Efficient for binary classification problems. **Cons:** Assumes a linear relationship between independent variables and log-odds. **Modeling equation:** $\log(p/(1-p)) = b_0 + b_1x_1 + ...$	```python\nfrom sklearn.linear_model import LogisticRegression\nmodel = LogisticRegression()\nmodel.fit(X, y)\n```

Associated functions commonly used

Function/Method Name	Brief Description	Code Syntax
train_test_split	Splits the dataset into training and testing subsets to evaluate the model's performance.	```python\nfrom sklearn.model_selection import train_test_split\nX_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)\n```
StandardScaler	Standardizes features by removing the mean and scaling to unit variance.	```python\nfrom sklearn.preprocessing import StandardScaler\nscaler = StandardScaler()\nX_scaled = scaler.fit_transform(X)\n```
log_loss	Calculates the logarithmic loss, a performance metric for classification models.	```python\nfrom sklearn.metrics import log_loss\nloss = log_loss(y_true, y_pred_proba)\n```
mean_absolute_error	Calculates the mean absolute error between actual and predicted values.	```python\nfrom sklearn.metrics import mean_absolute_error\nmae = mean_absolute_error(y_true, y_pred)\n```
mean_squared_error	Computes the mean squared error between actual and predicted values.	```python\nfrom sklearn.metrics import mean_squared_error\nmse = mean_squared_error(y_true, y_pred)\n```
root_mean_squared_error	Calculates the root mean squared error (RMSE), a commonly used metric for regression tasks.	```python\nfrom sklearn.metrics import mean_squared_error\nimport numpy as np\nrmse = np.sqrt(mean_squared_error(y_true, y_pred))\n```
r2_score	Computes the R-squared value, indicating how well the model explains the variability of the target variable.	```python\nfrom sklearn.metrics import r2_score\nr2 = r2_score(y_true, y_pred)\n```

Author(s)

Jeff Grossman

Abhishek Gagneja

Skills Network

IBM