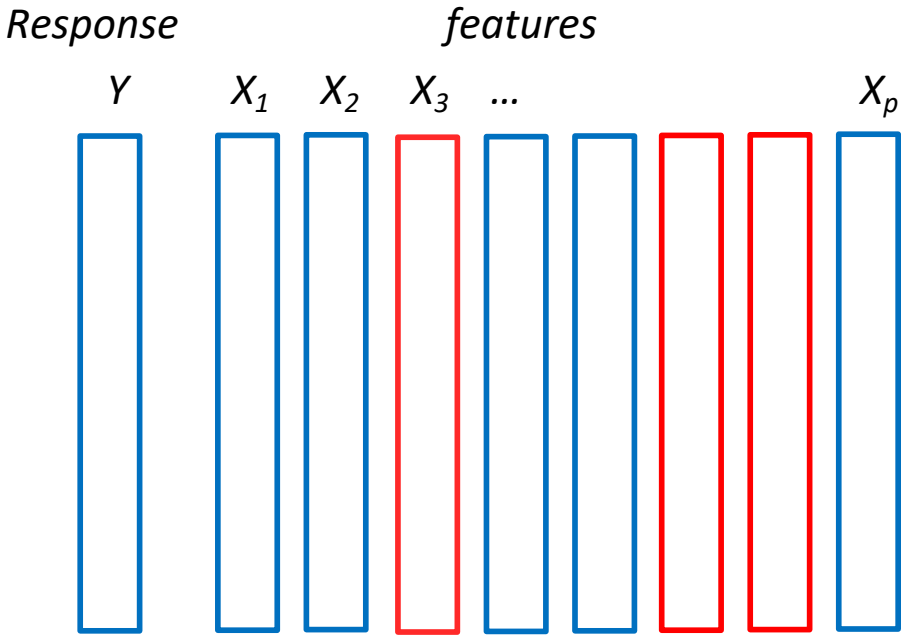
Regression vs Classification Learning Problems

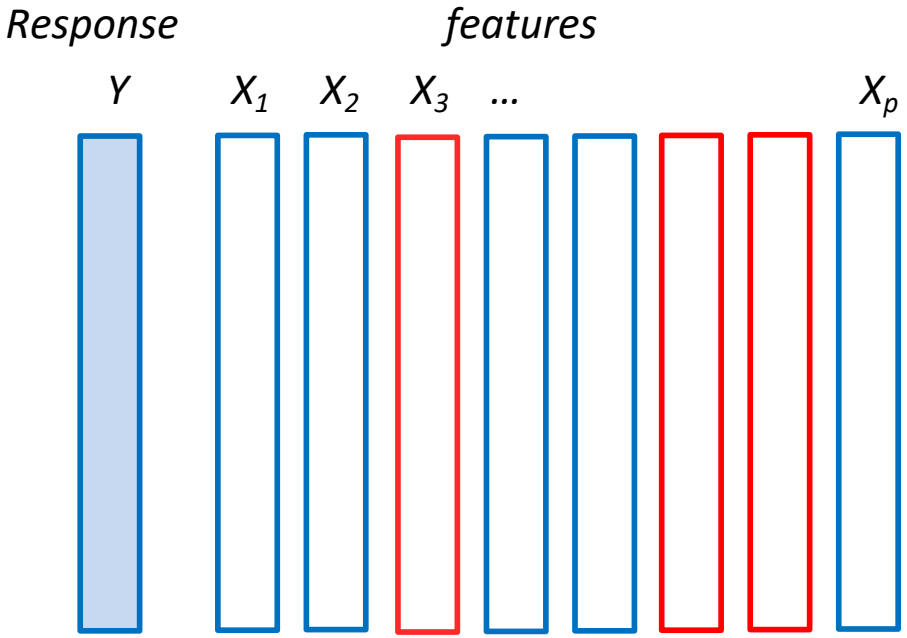
LINEAR REGRESSION MODELS



Blue for numeric variable

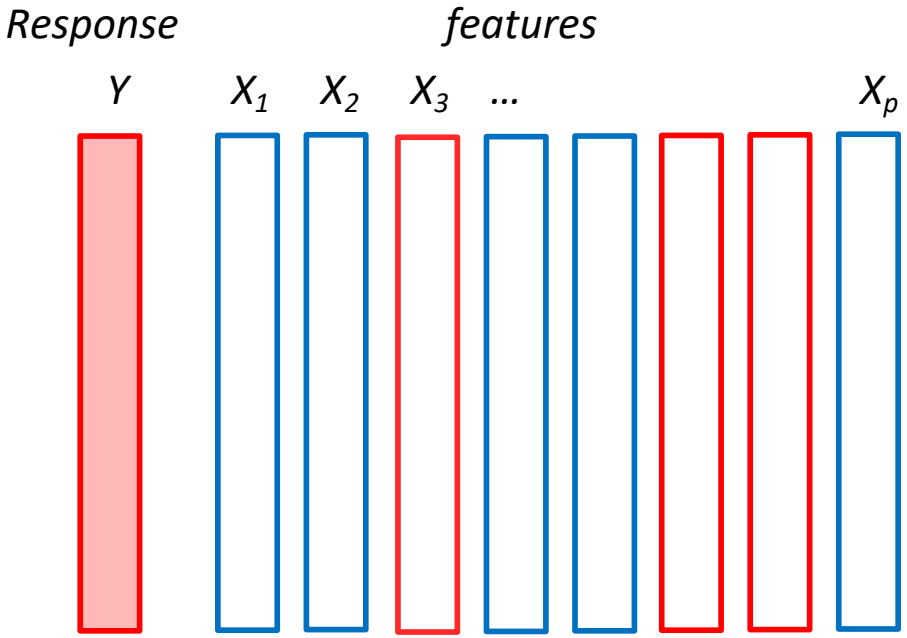
Red for categorical predictor

LINEAR REGRESSION MODELS



Response is numeric in Regression problems

CLASSIFICATION



Response is categorical in Classification problems

LINEAR REGRESSION MODELS

- *Response is numeric for Linear Regression models*

- *Standard Linear Regression models assume*
  the response is a normal r. variable

REGRESSION VS CLASSIFICATION

- *Regression models are evaluated on MSE, R^2, AIC, adj-R^2*
- *Classification models are evaluated on accuracy rate*

REGRESSION VS CLASSIFICATION

- *Regression models are evaluated on MSE, R^2, AIC, adj-R^2*
- *Classification models are evaluated on accuracy rate*

- *Accuracy rate is the number of correct predicted categories divided by total number of predictions*
- *For Cross Validation use the Accuracy Rate from the Test set (Test Accuracy Rate)*

Stratified Sampling for Classification Problems

Data set

20% (+) category Y=1

80% (-) category Y=0

training set

20% (+) category Y=1

80% (-) category Y=0

test set

20% (+) category Y=1

80% (-) category Y=0

HOLDOUT Cross Validation on Classification Problems

- Split the data into *train* and *test* sets. But make sure that the proportions of rows from each category are similar in the *train* and *test* sets, as in the dataset

```
X_train,X_test,y_train,y_test = train_test_split(X,y,
                                                 stratify=y,
                                                 random_state=66)
```

K-FOLD Cross Validation on Classification Problems

- Split the data into *k*-folds
- Make sure that the proportions of rows from each category are similar (across all folds), as they are in the whole dataset
- kfold = StratifiedKFold(n_splits)

```
from sklearn.model_selection import StratifiedKFold
from sklearn.model_selection import cross_val_score
```

```
kfold = StratifiedKFold(n_splits = 5,shuffle = True,random_state=1)
```

```
model1 = LogisticRegression(solver = 'lbfgs',max_iter = 10000)
scores = cross_val_score(model1,X,y,cv=kfold)
scores
```

Statistical Learning Models	Machine Learning Models
• Regression Models • Logistic Regression • Ridge, LASSO regression • Naive Bayes • Discriminant Analysis	• KNN • Decision Trees (CART) • Random Forest • Gradient Boosting • Support Vector Machines • Neural Networks

Statistical Learning Models	Machine Learning Models
• Regression Models • Logistic Regression • Ridge, LASSO regression • Naive Bayes • Discriminant Analysis	• KNN • Decision Trees (CART) • Random Forest • Gradient Boosting • Support Vector Machines • Neural Networks

Families of Machine Learning Models

for both classification and Regression problems

- KNNs
- CARTs
- SVMs
- NNs