Machine Learning Process #3

Lucas Sousa e Douglas Chielle

Huawei / IFCE

February 23, 2021

Agenda

1. Model Learning Performance Evaluation
   - Classification
   - Regression

2. Other Key Machine Learning Methods
   - Gradient Descent
   - Parameters and Hyperparameters in Models
   - Hyperparameter Search Procedure
   - Cross-validation

Agenda

1. Model Learning Performance Evaluation
   - Classification
   - Regression

2. Other Key Machine Learning Methods
   - Gradient Descent
   - Parameters and Hyperparameters in Models
   - Hyperparameter Search Procedure
   - Cross-validation

Agenda

1. Model Learning Performance Evaluation
   - Classification
   - Regression
2. Other Key Machine Learning Methods
   - Gradient Descent
   - Parameters and Hyperparameters in Models
   - Hyperparameter Search Procedure
   - Cross-validation

Classification

Classification

Receiving Operating Characteristic (ROC) curves

- Shows the sensitivity/specificity trade-off of a classifier for all possible classification thresholds

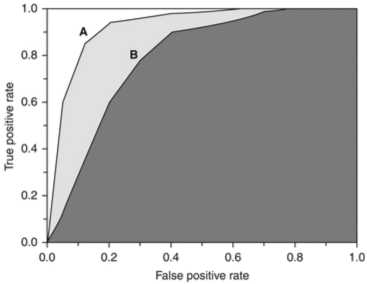
Classification

Receiving Operating Characteristic (ROC) curves

- Shows the sensitivity/specificity trade-off of a classifier for all possible classification thresholds

Area under a ROC curve

- Abbreviated as AUC, is a single scalar value that measures the overall performance of a binary classifier



Classification

Classification

Performance evaluation example

- We have trained a machine learning model to identify whether the object in an image is a cat. Now we use 200 pictures to verify the model performance. Among the 200 images, objects in 170 images are cats, while others are not. The identification result of the model is that objects in 160 images are cats, while others are not.

	yes	no
yes	140	30
no	20	10

Performance evaluation example

- We have trained a machine learning model to identify whether the object in an image is a cat. Now we use 200 pictures to verify the model performance. Among the 200 images, objects in 170 images are cats, while others are not. The identification result of the model is that objects in 160 images are cats, while others are not.

	yes	no
yes	140	30
no	20	10

- Accuracy rate?

Classification

Performance evaluation example

- We have trained a machine learning model to identify whether the object in an image is a cat. Now we use 200 pictures to verify the model performance. Among the 200 images, objects in 170 images are cats, while others are not. The identification result of the model is that objects in 160 images are cats, while others are not.

	yes	no
yes	140	30
no	20	10

- Accuracy rate? $\frac{150}{200} = 0.75$

Performance evaluation example

- We have trained a machine learning model to identify whether the object in an image is a cat. Now we use 200 pictures to verify the model performance. Among the 200 images, objects in 170 images are cats, while others are not. The identification result of the model is that objects in 160 images are cats, while others are not.

	yes	no
yes	140	30
no	20	10

- Accuracy rate? $\frac{150}{200} = 0.75$
- Precision?

Performance evaluation example

- We have trained a machine learning model to identify whether the object in an image is a cat. Now we use 200 pictures to verify the model performance. Among the 200 images, objects in 170 images are cats, while others are not. The identification result of the model is that objects in 160 images are cats, while others are not.

	yes	no
yes	140	30
no	20	10

- Accuracy rate? $\frac{150}{200} = 0.75$
- Precision? $\frac{140}{140+20} = 0.875$

Classification

Performance evaluation example

- We have trained a machine learning model to identify whether the object in an image is a cat. Now we use 200 pictures to verify the model performance. Among the 200 images, objects in 170 images are cats, while others are not. The identification result of the model is that objects in 160 images are cats, while others are not.

	yes	no
yes	140	30
no	20	10

- Accuracy rate? $\frac{150}{200} = 0.75$
- Precision? $\frac{140}{140+20} = 0.875$
- Recall?

Classification

Performance evaluation example

- We have trained a machine learning model to identify whether the object in an image is a cat. Now we use 200 pictures to verify the model performance. Among the 200 images, objects in 170 images are cats, while others are not. The identification result of the model is that objects in 160 images are cats, while others are not.

	yes	no
yes	140	30
no	20	10

- Accuracy rate? $\frac{150}{200} = 0.75$
- Precision? $\frac{140}{140+20} = 0.875$
- Recall? $\frac{140}{140+30} = 0.824$

Agenda

Regression

Regression

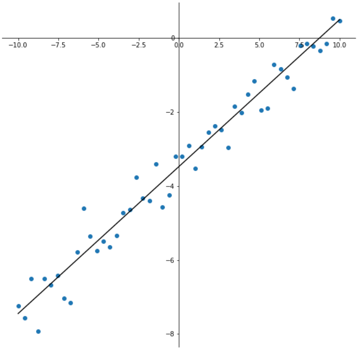
- A method for fitting a curve (not necessarily a straight line) through a set of points using some goodness-of-fit criterion.

Regression

- A method for fitting a curve (not necessarily a straight line) through a set of points using some goodness-of-fit criterion.
- The most common type of regression is linear regression.

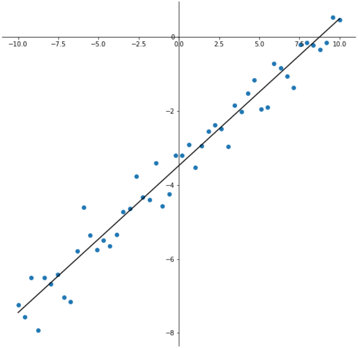
Regression

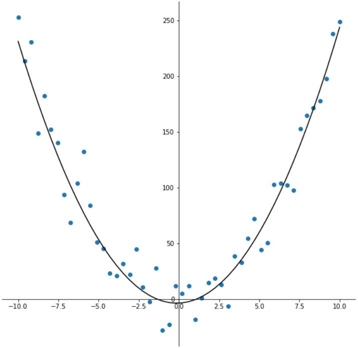
- A method for fitting a curve (not necessarily a straight line) through a set of points using some goodness-of-fit criterion.
- The most common type of regression is linear regression.



Regression

- A method for fitting a curve (not necessarily a straight line) through a set of points using some goodness-of-fit criterion.
- The most common type of regression is linear regression.





Regression metrics

Regression metrics

Mean Squared Error (MSE)	$\mathrm{MSE} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2$

Regression metrics

Mean Squared Error (MSE)	$\mathrm{MSE} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2$
Mean Absolute Error (MSE)	$\mathrm{MAE} = \frac{1}{N}\sum_{n=1}^{N}\lvert y_n - \hat{y}_n\rvert$

Regression

Regression metrics

Mean Squared Error (MSE)	$\mathrm{MSE} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2$
Mean Absolute Error (MSE)	$\mathrm{MAE} = \frac{1}{N}\sum_{n=1}^{N}\lvert y_n - \hat{y}_n\rvert$
R-squared (R^2)	$R^2 = 1 - \frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{\sum_{n=1}^{N}(y_n - \bar{y}_n)^2}$

Agenda

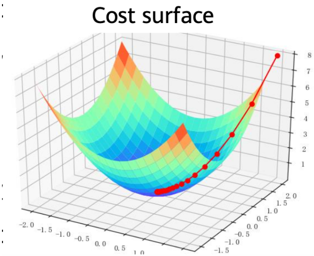
1. Model Learning Performance Evaluation
   - Classification
   - Regression
2. Other Key Machine Learning Methods
   - Gradient Descent
   - Parameters and Hyperparameters in Models
   - Hyperparameter Search Procedure
   - Cross-validation

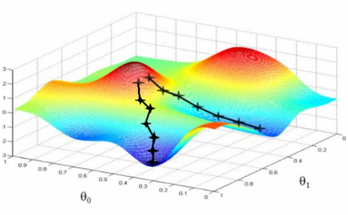
Agenda

1. Model Learning Performance Evaluation
   - Classification
   - Regression
2. Other Key Machine Learning Methods
   - Gradient Descent
   - Parameters and Hyperparameters in Models
   - Hyperparameter Search Procedure
   - Cross-validation

Gradient Descent

Gradient Descent





Gradient Descent

- The gradient descent method uses the negative gradient direction of the current position as the search direction, which is the steepest direction. The formula is as follows

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \nabla f_{\boldsymbol{w}_t}(\boldsymbol{x}) \tag{1}$$

Gradient Descent

- The gradient descent method uses the negative gradient direction of the current position as the search direction, which is the steepest direction. The formula is as follows

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \nabla f_{\boldsymbol{w}_t}(\boldsymbol{x}) \tag{1}$$

- η indicates the learning rate

Gradient Descent

- The gradient descent method uses the negative gradient direction of the current position as the search direction, which is the steepest direction. The formula is as follows

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \nabla f_{\boldsymbol{w}_t}(\boldsymbol{x}) \tag{1}$$

- η indicates the learning rate
- The value of the objective function changes very little, or the maximum number of iterations is reached.

Batch Gradient Descendent

- Batch Gradient Descent (BGD) uses the samples (m in total) in all datasets to update the weight parameter based on the gradient value at the current point.

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \frac{1}{m} \sum_{i=1}^{m} \nabla f_{\boldsymbol{w}_t}(\boldsymbol{x_i}) \tag{2}$$

Stochastic Gradient Descent

- Stochastic Gradient Descent (SGD) randomly selects a sample in a dataset to update the weight parameter based on the gradient value at the current point.

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \nabla f_{\boldsymbol{w}_t}(\boldsymbol{x}) \tag{3}$$

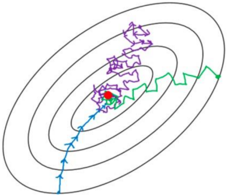
Mini-Batch Gradient Descent

- Mini-Batch Gradient Descent (MBGD) combines the features of BGD and SGD and selects the gradients of n samples in a dataset to update the weight parameter.

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \frac{1}{n} \sum_{i=k}^{t+n-1} \nabla f_{\boldsymbol{w}_t}(\boldsymbol{x_i}) \tag{4}$$

Gradient Descent Comparison

Gradient Descent Comparison



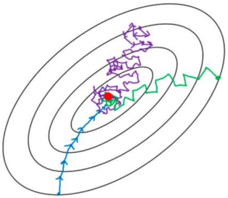
BGD
Uses **all** training samples for training each time.

SGD
Uses **one** training sample for training each time.

MBGD
Uses a certain number of training samples for training each time.

- In the SGD, samples selected for each training are stochastic. Such instability causes the loss function to be unstable or even causes reverse displacement when the loss function decreases to the lowest point.

Gradient Descent Comparison



BGD
Uses **all** training samples for training each time.

SGD
Uses **one** training sample for training each time.

MBGD
Uses a certain number of training samples for training each time.

- BGD has the highest stability but consumes too many computing resources. MBGD is a method that balances SGD and BGD.

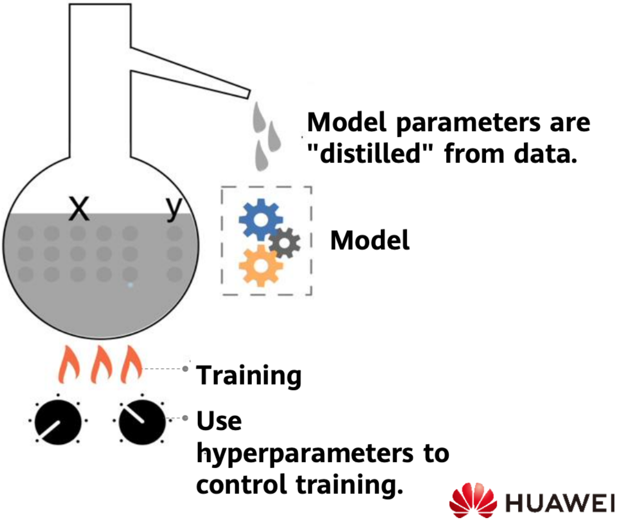
Agenda

Parameters and Hyperparameters in Models

Parameters and Hyperparameters in Models

- The model contains not only parameters but also hyperparameters. The purpose is to enable the model to learn the optimal parameters.
  - Parameters are automatically learned by models.
  - Hyperparameters are manually set.

Gradient Descent Comparison



Hyperparameters of a Model

- Model hyperparameters are external configurations of models.
    - Often used in model parameter estimation process
    - Often specified by the practitioner
    - Can often be set using heuristics
    - Often tuned for a given predictive modeling problem

Hyperparameters of a Model

- Common model hyperparameters
  - λ during Lasso/Ridge regression
  - Learning rate for training a neural network, number of iterations, batch size, activation function, and number of neurons
  - C and σ in support vector machines (SVM)
  - K in k-nearest neighbor (KNN)
  - Number of trees in a random forest

Agenda

Hyperparameter Search Procedure

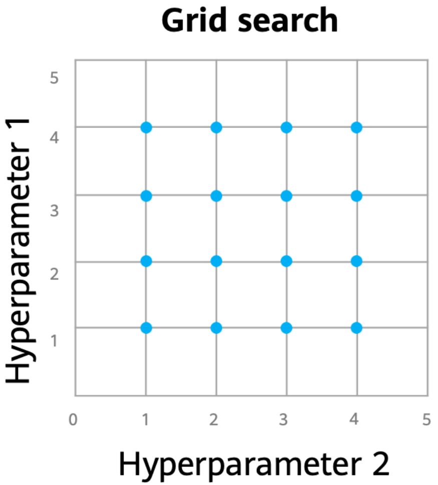
Procedure for searching hyperparameters

1. Dividing a dataset into a training set, validation set, and test set.
2. Optimizing the model parameters using the training set based on the model performance indicators.
3. Searching for the model hyper-parameters using the validation set based on the model performance indicators.
4. Perform step 2 and step 3 alternately. Finally, determine the model parameters and hyperparameters and assess the model using the test set.

Search algorithm (step 3)

- **Grid search**
- **Random search**
- Heuristic intelligent search
- Bayesian search

Hyperparameter Searching Method - Grid Search



Hyperparameter Searching Method - Grid Search

Hyperparameter Searching Method - Grid Search

- Grid search attempts to **exhaustively search** all possible hyperparameter combinations to form a hyperparameter value grid.

Hyperparameter Searching Method - Grid Search

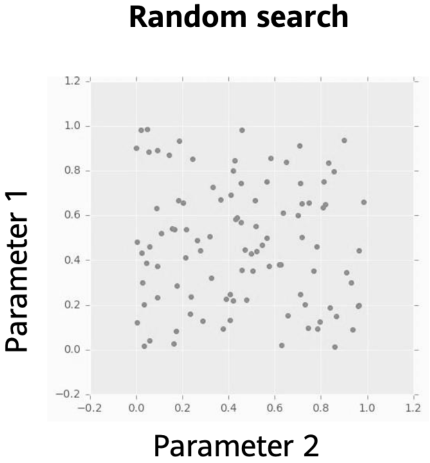
- Grid search attempts to **exhaustively search** all possible hyperparameter combinations to form a hyperparameter value grid.
- In practice, the range of hyperparameter values to search is specified manually.

Hyperparameter Searching Method - Grid Search

Hyperparameter Searching Method - Grid Search

- This method works well when the number of hyperparameters is relatively small. Therefore, it is applicable to generally machine learning algorithms but inapplicable to neural networks

Hyperparameter Searching Method - Random Search



Hyperparameter Searching Method - Random Search

Hyperparameter Searching Method - Random Search

- When the hyperparameter search space is large, random search is better than grid search.

Hyperparameter Searching Method - Random Search

- When the hyperparameter search space is large, random search is better than grid search.
- In random search, each setting is sampled from the distribution of possible parameter values, in an attempt to find the best subset of hyperparameters.

Hyperparameter Searching Method - Random Search

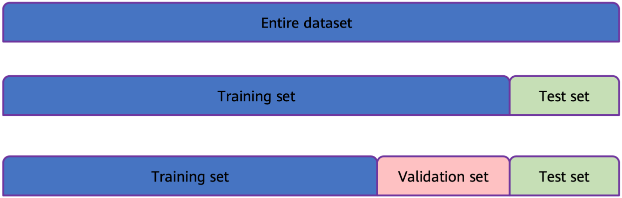
- When the hyperparameter search space is large, random search is better than grid search.
- In random search, each setting is sampled from the distribution of possible parameter values, in an attempt to find the best subset of hyperparameters.
- Note:
  - Search is performed within a coarse range, which then will be narrowed based on where the best result appears.
  - Some hyperparameters are more important than others, and the search deviation will be affected during random search.

Agenda

Cross-validation

Cross-validation

- It is a statistical analysis method used to validate the performance of a classifier. The basic idea is to divide the original dataset into two parts: training set and validation set. Train the classifier using the training set and test the model using the validation set to check the classifier performance.



K-fold Cross-validation

K-fold Cross-validation

- Divide the raw data into k groups (generally, evenly divided).

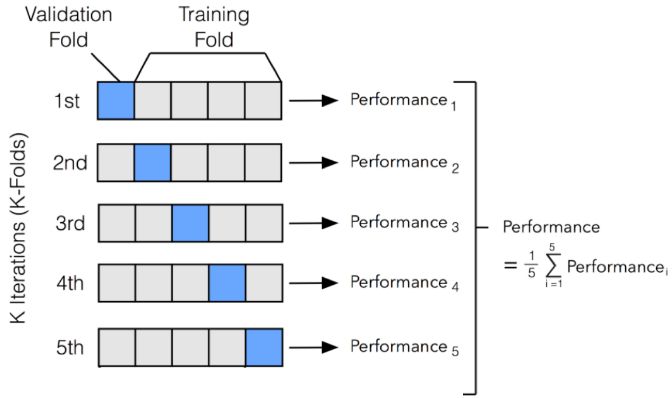
K-fold Cross-validation

- Divide the raw data into k groups (generally, evenly divided).
- Use each subset as a validation set, and use the other $k1$ subsets as the training set. A total of k models can be obtained.

K-fold Cross-validation

- Divide the raw data into k groups (generally, evenly divided).
- Use each subset as a validation set, and use the other $k1$ subsets as the training set. A total of k models can be obtained.
- Use the mean classification accuracy of the final validation sets of k models as the performance indicator of the k-fold classifier.

k-fold Cross-validation



Thanks for your attention