|  |  |
| --- | --- |
| **Supervised Learning** | |
| **Linear Regression:** *Trend, Market estimates, Forecasts* | |
| **Step 1. Hypothesis:**  **Step 2. Cost**  **Step 3: Gradients** | A close up of a map  Description automatically generated |
| A close up of a map  Description automatically generated |
| **Logistic Regression:** *Binary classes* | |
| **Step 1. Hypothesis:**  **Step 2. Cost**  **Step 3. Gradients:** | A picture containing food  Description automatically generated |
| **Support Vector Machines (SVM)** | |
| **Cost** | Related image |
| **SVM with Gaussian Kernel** | |
| **Step 1. Hypothesis** *Given x, compute features , parameters*  *Predict “y=1” if*  **Step 2. Training**  *min* | A picture containing text, map  Description automatically generated |
|  | |
| **Neural network:** *Pattern recognition, Fraud detection, Deep learning.* | |
| **Step 1. Randomly initialize weights**  Initialize parameters (i.e. )  **Step 2. Forward propagation**      **Step 3. Cost function**  **Step 4. Backpropagation to compute partial derivatives**  “error” of node in layer .   |  |  | | --- | --- | |  |  | |  |   **Step 5.** Use gradient checking to compare computed using backpropagation vs. using numerical estimate of gradient of .  **Step 6.** Use gradient descent or advanced optimization method with backpropagation to try to minimize as a function of parameters .  result = minimize(cost\_func, initial\_nn\_params, method='CG', jac=grad\_func,  options={'disp': True, 'maxiter': 50.0})  nn\_params = result.x  Jcost = result.fun | A picture containing building, window, drawing  Description automatically generated |

|  |  |
| --- | --- |
| **Unsupervised Learning** | |
| **K-means:** *Clustering* | |
| **Step 1. Centroids**  denotes the index of cluster centroids closet to  **Step 2. Means**  denotes the average(mean) of points assigned to cluster k  **Step 3. Cost function** | K=3  A close up of a map  Description automatically generated |
| **Principal Component Analysis** (PCA):  *Dimensionality Reduction, Facial recognition, Data compression, Computer vision* | |
| **Step1. Feature scaling (Mean normalization)**  Mean:  Standard deviation:  Mean normalize:  **Step 2. Calculate U, S, V.**    U, S, V = numpy.linalg.svd(sigma)  Ureduce = U[:, 0:K].T  Z = Ureduce\*X = X\_norm \* U[:, 0:K]  X\_approximate = X\_recovered = Z \* U[:, 0:K].T  **Step 3. Pick the smallest value of k,**  99% of variance retained | PC Scatterplot |
| **Anomaly Detection:** *Fraud detection, Intrusion detection, system health, monitoring* | |
| **Gaussian (Normal) distribution**  **Mean:**  **Variance:**  **Probability:** | A close up of a map  Description automatically generated |
| **Original model** | |
| **Step 1. Choose feature**  Training set:  Density estimation: **1, 2, …, n**  Choose features that might be indicative of anomalous examples.  **Step 2. Fit parameters**  **Step 3. Given new example , compute**  **Probability** | A close up of a piece of paper  Description automatically generated |
| **Multivariate Gaussian** | |
| **Step 1. Choose feature**  Training set:  Density estimation: 1, 2, …, n  **Step 2. Fit parameters**  Parameters: (covariance matrix)  Mean:  Variance:  **Step 3. Given new example , compute**  Diagonal Sigma:  Probability: | A close up of text on a white background  Description automatically generated |

Machine Learning

Allen Sun

Contents

[1 Machine learning 2](#_Toc31926201)

[2 Supervised learning 2](#_Toc31926202)

[2.1 Linear regression 2](#_Toc31926203)

[2.2 Logistic regression 4](#_Toc31926204)

[2.3 Support vector machine (SVM) 5](#_Toc31926205)

[2.3.1 No kernel (“linear kernel”) 6](#_Toc31926206)

[2.3.2 SVM with kernels, also called gaussian kernel 6](#_Toc31926207)

[2.4 Neural Network 7](#_Toc31926208)

[3 Unsupervised learning 9](#_Toc31926209)

[3.1 K-means 9](#_Toc31926210)

[3.2 Principal component analysis (PCA) 11](#_Toc31926211)

[3.3 Anomaly detection 12](#_Toc31926212)

[3.3.1 Original model 12](#_Toc31926213)

[3.3.2 Multivariate Gaussian 13](#_Toc31926214)

[3.3.3 Collaborative filtering 14](#_Toc31926215)

[4 Tips 16](#_Toc31926216)

[4.1 Diagnostics 16](#_Toc31926217)

[4.2 Model Selection and Train/Validation/Test Sets 16](#_Toc31926218)

[4.3 Gaussian (Normal) distribution 17](#_Toc31926219)

[4.4 F1-score 17](#_Toc31926220)

[4.5 Greek Alphabet 18](#_Toc31926221)

# Machine learning

Machine learning is the science of getting computers to learn, without being explicitly programmed.

* **Cost function** – The cost function J is commonly used to assess the performance of a model with parameters and hypothesis function .
* This minimization objective is expressed using the following notation, which simply states that we want to find the  which minimizes the cost .

A picture containing screenshot

Description automatically generated

Figure 1 Change of cost function J with iterations

* **Gradient descent –** The gradient descent is used to minimize cost function J with learning rate and derivative .

A close up of a map

Description automatically generated

Figure 2 Example of batch gradient descent

# Supervised learning

Already know what our correct output should look like, having the idea that there is a relationship between the labeled input and the labeled output.

## Linear regression

**Motivation**: Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).

**Key words.** continuous, individual

1. Data & Hypothesis

Input data: X (m, n); dataset: ; features: ;

parameters: (n,1); actual output: y(m,1).

, ,

1. Hypothesis:
2. Cost function:

We can measure the accuracy of our hypothesis function by using a **cost function**. This takes an average difference (actually a fancier version of an average) of all the results of the hypothesis with inputs from x's and the actual output y's.

Minimize cost:

1. Gradients

A close up of a map

Description automatically generated

Figure 3 Linear regression descent

When specifically applied to the case of linear regression, a new form of the gradient descent equation can be derived. We can substitute our actual cost function and our actual hypothesis function and modify the equation to:

Rate: , regulation:

Repeat until convergence: {

}

 <1. Intuitively you can see it as reducing the value of  by some amount on every update.

* **Advice:** Debugging a learning algorithm:

Suppose you have implemented regularized linear regression to predict housing prices. However, when you test your hypothesis on a new set of houses, you find that it makes unacceptably large errors in its predictions. What should you try next?

* Get more training examples
* Try smaller sets of features
* Try getting additional features
* Try adding polynomial features
* Try decreasing
* Try increasing

## Logistic regression

**Motivation**: the logistic regression is used to model the probability of a certain class or event existing.

**Key words:** classification, discrete, collective

1. Sigmoid function

In order to get our discrete 0 or 1 classification, we can translate the output of the hypothesis function as follows:

The way our logistic function g behaves is that when its input is greater than or equal to zero, its output is greater than or equal to 0.5:

The following image shows us what the sigmoid function looks like:

A picture containing kitchen, white, light, large

Description automatically generated

If our input to g is , then that means:

From these statements we can now say:

1. Cost function:

Minimize cost:

1. Gradients descent:

A close up of a map

Description automatically generated

Figure 4 Logistic gradients descent

when computing the equation, we should continuously update the two following equations:

Repeat: {

}

## Support vector machine (SVM)

**Motivation**: SVMs can efficiently perform a non-linear classification implicitly mapping their inputs into high-dimensional feature spaces.

**Key words:** Mapping, discrete, collective

A picture containing drawing

Description automatically generated

Figure 5 SVM Kernel mapping

### No kernel (“linear kernel”)

Predict “y = 1” if

1. Hypothesis:
2. Cost function:

A picture containing object, clock

Description automatically generated

Large C: Lower bias, high variance.

Small C: Higher bias, low variance.

### SVM with kernels, also called gaussian kernel

#### Hypothesis

*Given x, compute features , parameters*

*Predict “y=1” if*

1. Training

*min*

|  |  |
| --- | --- |
| Large : Features vary more smoothly.  Higher bias, lower variance. | A picture containing clock  Description automatically generated |
| Small : Features vary less smoothly.  Lower bias, higher variance. | A close up of a logo  Description automatically generated |

1. Multiclass classification

Many SVM packages already have built-in multiclass classification functionality.

Otherwise, use one-vs.-all method. (Train SVMs, one to distinguish from the rest, for ), get Pick class with largest

A screenshot of a cell phone

Description automatically generated

## Neural Network

**Motivation:** behaving like a human brain.

**Key words:** classification, layers, collective.

Non-linear Classification

|  |  |
| --- | --- |
| A picture containing building, window, drawing  Description automatically generated | no. of units (not counting bias unit) in layer |

|  |  |
| --- | --- |
| Binary classification  1 output unit | Multi-class classification (K classes)  E.g.  K output units |

* Training a neural network (e.g. L=4)
* Pick a network architecture (connectivity pattern between neurons)
* No. of input units: Dimension of features
* No. output units: Number of classes
* Reasonable default: 1 hidden layer, or if >1 hidden layer, have same no. of hidden units in every layer (usually the more the better)

1. Randomly initialize weights

Initialize parameters

Initialize each to a random value in (i.e. )

epsilon = 0.12

theta = np.random.rand(input\_layer, output\_layer) \* 2\*epsilon - epsilon

1. Implement forward propagation to get for any

|  |  |  |  |
| --- | --- | --- | --- |
| A picture containing building, window, drawing  Description automatically generated | | | |
| Input layer | Hidden layers | | Output layer |
|  |  |  |  |

1. Implement code to compute cost function
2. Implement backpropagation to compute partial derivatives

“error” of node in layer .

|  |  |  |  |
| --- | --- | --- | --- |
| A picture containing building, window, drawing  Description automatically generated | | | |
| Input layer | Hidden layers | | Output layer |
|  |  |  |  |

Perform forward propagation and backpropagation using example   
(Get activations and delta terms for ).

* Set (from all, ).
* Set
* Perform forward propagation to compute for
* Using y, compute
* Compute

|  |  |
| --- | --- |
|  |  |
|  |

1. Use gradient checking to compare computed using backpropagation vs. using numerical estimate of gradient of .
2. Implement backpropagation to compute DELTA VECTOR (unrolled ).
3. Implement numerical gradient check to compute **gradient approximation**.
4. Make sure they give similar values.

diff = slin.norm(numgrad-grad)/slin.norm(numgrad+grad)

print('If your backpropagation implementation is correct, then \n\

the relative difference will be small (less than 1e-9). \n\

\nRelative Difference: ', diff)

1. Turn off gradient checking. Using backprop code for learning.

* **Important:** Be sure to disable your gradient checking code before training your classifier. If you run numerical gradient computation on every iteration of gradient descent (or in the inner loop of cost function your code will be very slow).

1. Use gradient descent or advanced optimization method with backpropagation to try to minimize as a function of parameters .

result = minimize(cost\_func, initial\_nn\_params, method='CG', jac=grad\_func,

options={'disp': True, 'maxiter': 50.0})

nn\_params = result.x

Jcost = result.fun

# Unsupervised learning

Fund hidden pattern in unlabeled data with little or no idea what our results should look like.

Don't necessarily know the effect of the variables.

## K-means

**Motivation:** k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

**Key words:** clustering

1. Choose the value of K

Sometimes, you’re running K-means to get clusters to use for some later/downstream purpose. Evaluate K-means based on a metric for how well it performs for that later purpose.

A close up of text on a white background

Description automatically generated

Figure 6 T-shirt size example

1. Initialize centroids

Random initialize K clustering centroids

For i = 1 to 100{

Randomly pick examples from given points as K () different centroids

initial\_centroids = random.sample(X.tolist(), K)

cost function

*Return centroids of the smallest J.*

**}**

1. Iteration

Set Iterate times, max\_iters

For iters = 1 to max\_iters {

denotes the index of cluster centroids closet to .

denotes the average(mean) of points assigned to cluster k.

}

A close up of a map

Description automatically generated

Figure 7 K clusters

## Principal component analysis (PCA)

**Motivation:** Principal component analysis is a statistical procedure to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

**Key words:** Dimensionality reduction, covariance

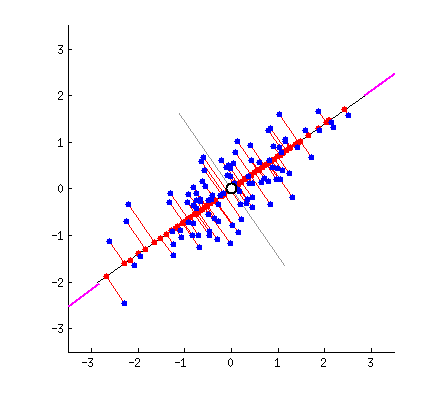


Figure 8 Reduce 2 dimensions to 1 dimension

1. **Data processing**

Training set:

Processing: feature scaling (mean normalization) to ensure every feature has zero mean.

1. Mean value:
2. Standard deviation
3. Replace each
4. PCA algorithm.

Reduce data from n-dimensions to k-dimensions

Principal Component Analysis (PCA) algorithm:

U, S, V = numpy.linalg.svd(sigma)

Dimension matrix:

Ureduce = U[:, 0:K].T

Reduced K-dimensions data:

Z = Ureduce \* X = X\_norm \* U[:, 0:K]

Recover data to n-dimentions:

X\_approximate = X\_recovered = Z \* U[:, 0:K].T

1. **Choosing K** (number of principal components)**:**

Pick the smallest value of K.

**Check if**

**99% of variance retained.**

**A screenshot of a cell phone

Description automatically generated**

## Anomaly detection

**Motivation:** anomaly detection is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data.

**Key words:** anomaly, filtering.

### Original model

1. Choose feature:

Training set:

Density estimation: **1, 2, …, n**

Choose features that might be indicative of anomalous examples.

1. Fit parameters:
2. Given new example , compute

Flag an anomaly if

### Multivariate Gaussian

Don’t model etc. separately. Model all in one go.

1. Choose feature

Training set:

Density estimation: **1, 2, …, n**

1. Fit parameters

Parameters: (covariance matrix)

1. Given new example , compute

Flag an anomaly if

* The importance of real-number evaluation

1. When developing a learning algorithm (choosing features, etc.), making decisions is much easier if we have a way of evaluating our learning algorithm.

Assume we have some labeled data, of anomalous and non-anomalous examples. ( if normal, if anomalous).

Training set 60%: (assume normal examples/not anomalous)

Cross validation set 20%:

Test set 20%:

1. Algorithm evaluation

Possible evaluation metrics

Precision

Recall

F1-Score

* **Comparison between these two models**

**A screenshot of a cell phone

Description automatically generated**

### Collaborative filtering

Recommender system

1. Problem formulation

if user has rated movie (0 otherwise)

rating by user on movie (if defined)

= parameter vector for user

= feature vector for movie   
For user , movie, predicted rating:

= no. of movies rated by user

1. Initialization

Initialize and to small random values.

1. Cost function
2. Minimize and

Min

1. using gradient descent (or an advanced optimization algorithm). E.g. for every :

Gradients

1. Prediction.

For a user with parameters and a movie with(learned) features , predict a star rating of

is the mean value of rated users for movie .

A screenshot of a cell phone

Description automatically generated

# Tips

## Diagnostics

* **Bias** – The bias of a model is the difference between the expected prediction and the correct model that we try to predict for given data points.
* **Variance** – The variance of a model is the variability of the model prediction for given data points.
* **Bias/variance tradeoff** – The simpler the model, the higher the bias, and the more complex the model, the higher the variance.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Underfitting | Just right | Overfitting |
| Symptoms | - High training error  - Training error close to test error  - High bias | - Training error  - Slightly lower than test error | - Low training error  - Training error much lower than test error  - High variance |
| Regression |  |  |  |
| Classification |  |  |  |
| Deep Learning  (Neural network) |  |  |  |
| Remedies | - Complexify model  - Add more features  - Train longer |  | - Regularize Remedies  - Get more data |

## Model Selection and Train/Validation/Test Sets

One way to break down our dataset into the three sets is:

1. Training set: 60%
2. Cross validation set: 20%
3. Test set: 20%

We can now calculate three separate error values for the three different sets using the following method:

1. Optimize the parameters in Θ using the training set for each polynomial degree.
2. Find the polynomial degree d with the least error using the cross-validation set.
3. Estimate the generalization error using the test set with Jtest(Θ(d)), (d = theta from polynomial with lower error);

This way, the degree of the polynomial d has not been trained using the test set.

## Gaussian (Normal) distribution

A close up of a map

Description automatically generated

Figure 9 68% of the data is within 1 standard deviation, 95% is within 2 standard deviation, 99.7% is within 3 standard deviations

## F1-score

 F-score is a measure of a test's accuracy.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual class | |
|  |  | 1/+ | 0/- |
| Predicted  class | 1/+ | True positive  (TP) | False positive  (FP) |
| 0/- | False negative  (FN) | True negative  (TN) |

**Precision**

**Recall**

**F1-score**

## Greek Alphabet

A screenshot of a cell phone

Description automatically generated