

An Experience with Text Classification in *Datadays 2019*

Majid Hajiheidari Amirmohammad Asadi

April, 2019

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

TF-idf Vectorizer

Count Vectorizer

Embedding

Classification
Algorithms

Naive Bayes

CNN

Linear SVM

Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning

Comparison

The End

Divar Posts Dataset

- ▶ Released for DataDays 2019
- ▶ One million posts

بارگشت همه آگهی ها / املاک / اجاره مسکونی (آپارتمان، خانه، زمین) / آپارتمان



190متر/4خواب/فول محدوده کاج
دقایقی پیش

نشان کردن

شروع چت

دریافت اطلاعات تماس

آپارتمان	دسته بندی
تهران سعادت آباد	محل
ارائه	نوع آگهی
شخصی	آگهی دهنده
چهار	تعداد اتاق
۱۹۰	متراژ

بارگشت همه آگهی ها / سرگرمی و فراغت / دوجرخه/اسکیت/اسکوتر

بارگشت همه آگهی ها / سرگرمی و فراغت / دوجرخه/اسکیت/اسکوتر



دوجرخه مریدا 7-300 سال ۲۰۱۷
۲ ساعت پیش

نشان کردن

شروع چت

دریافت اطلاعات تماس

دوجرخه/اسکیت/اسکوتر	دسته بندی
تهران میدان آزادی	محل
فروشی	نوع آگهی
۵/۸۰۰۰۰۰۰ تومان	قیمت

با سلام یک دستگاه دوجرخه مریدا 7-300 سال ۲۰۱۷ در حد اک اک سایز 27/5 تنه 18/5 با کمک باد ست اورزم دنده=طرق و

An Experience
with Text
Classification in
DataDays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

TF-idf Vectorizer
Count Vectorizer
Embedding

Classification
Algorithms

Naive Bayes
CNN
Linear SVM
Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning
Comparison

The End

Columns

- ▶ id
- ▶ archive_by_user
- ▶ published_at
- ▶ **cat1**
- ▶ **cat2**
- ▶ **cat3**
- ▶ city
- ▶ **title**
- ▶ **desc**
- ▶ price
- ▶ image_count
- ▶ platform
- ▶ mileage
- ▶ brand
- ▶ year
- ▶ type

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer
Count Vectorizer
Embedding

Classification
Algorithms

Naive Bayes
CNN
Linear SVM
Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning
Comparison

The End

The Problem: Categorization

- ▶ We need to categorize posts based on other posts features;
- ▶ We only use text features(title & description)!

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer
Count Vectorizer
Embedding

Classification
Algorithms

Naive Bayes
CNN
Linear SVM
Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning
Comparison

The End

Features

Cat3	Cat2	Cat1	Desc	Title
fridge-and-freezer	utensils-and-appliances	for-the-home	یخچال ارج کاملا سالم	یخچال ارج
nan	childrens-clothing-and-shoe	personal	دونه ای 28 سن 7 تا 9 تقریبا رنگ مناسب دختر و پسر میباشد مقطووووع پیامک پاسخگو نیستم	تعدادی کا پشن درحدنو
stereo-surround	audio-video	electronic-devices	سالم و با صدای فوق العاده قوی و با کیفیت. AUX رم و فلش هم میخوره به ضبط دو تیکه LG آمپلی دار هم دارم که داخله عکس مشخصه اونم تقدیم میکنم. یا علی	سینما خانگی
light	cars	vehicles	همه امکانات رو داره	خودرو پژو ۴۰۵
mobile-phones	mobile-tablet	Electronic-devices	بدون ضربه خوردگی و تعمیر	ایفون 6 گری ۶۴ گیگ

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

TF-idf Vectorizer

Count Vectorizer

Embedding

Classification
Algorithms

Naive Bayes

CNN

Linear SVM

Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning

Comparison

The End

No. of Classes

- ▶ We concatenate three category columns into one; for example:

cat1	cat2	cat3	concatenate
vehicles	cars	light	vehicles::cars::light

- ▶ Then, we have 87 unique combinations of categories, eg. 87 classes in our classification task.

Feature Extraction

Feature extraction is a dimensionality reduction process, where an initial set of raw variables is reduced to more manageable groups (features) for processing, while still accurately and completely describing the original data set.

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer
Count Vectorizer
Embedding

Classification
Algorithms

Naive Bayes
CNN
Linear SVM
Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning
Comparison

The End

Tf-idf Vectoizer

- ▶ Tf-idf stands for term frequency-inverse document frequency
- ▶ a statistical measure used to evaluate how important a word is to a document in a collection or corpus
- ▶ the tf-idf weight is composed by two terms:

TF Term Frequency, which measures how frequently a term occurs in a document.

$$TF(t) = \frac{\text{Number of times term } t \text{ appears in a document}}{\text{Total number of terms in the document}}$$

IDF Inverse Document Frequency, which measures how important a term is

$$IDF(t) = \ln \frac{\text{Total number of documents}}{\text{Number of documents with term } t \text{ in it}}$$

Tf-idf Vectorizer: An Example

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer
Count Vectorizer
Embedding

Classification
Algorithms

Naive Bayes
CNN
Linear SVM
Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning
Comparison

The End

Consider a document containing 100 words wherein the word *cat* appears 3 times. The term frequency (i.e., *tf*) for *cat* is then $tf(cat) = \frac{3}{100} = 0.03$. Now, assume we have 10 million documents and the word *cat* appears in one thousand of these. Then, the inverse document frequency (i.e., *idf*) is calculated as $idf(cat) = \ln \frac{10,000,000}{1,000} = 4$. Thus, the Tf-idf weight is the product of these quantities: $tf-idf(cat) = 0.03 * 4 = 0.12$.

Vectorizing the Text: Count Vectorizer

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

An example: We want to vectorize these 4 sentences¹:

1. Hello, how are you!
2. Win money, win from home.
3. Call me now
4. Hello, Call you tomorrow?

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer

Count Vectorizer

Embedding

Classification
Algorithms

Naive Bayes

CNN

Linear SVM

Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning

Comparison

The End

¹Example from Rahul Vasaikar

Vectorizing the Text: Count Vectorizer

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

1. We first build a vocabulary:

vocabulary =

{are, call, from, hello, home, how, me, money, now, tomorrow, win, you}

2. Then, we vectorize each sentence based on the occurance of each word:

	are	call	from	hello	home	how	me	money	now	tom...	win	you
1	1	0	0	1	0	1	0	0	0	0	0	1
2	0	0	1	0	1	0	0	1	0	0	2	0
3	0	1	0	0	0	0	1	0	1	0	0	0
4	0	1	0	1	0	0	0	0	0	1	0	1

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer

Count Vectorizer

Embedding

Classification
Algorithms

Naive Bayes

CNN

Linear SVM

Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning

Comparison

The End

Word Embedding

... when the input to a neural network contains symbolic categorical features (e.g. features that take one of k distinct symbols, such as words from a closed vocabulary), it is common to associate each possible feature value (i.e., each word in the vocabulary) with a d -dimensional vector for some d . These vectors are then considered parameters of the model, and are trained jointly with the other parameters.

— Page 49, Neural Network Methods in Natural Language Processing, 2017.

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer

Count Vectorizer

Embedding

Classification
Algorithms

Naive Bayes

CNN

Linear SVM

Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning

Comparison

The End

Word Embedding

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

It requires that document text be cleaned and prepared such that each word is one-hot encoded. The size of the vector space is specified as part of the model, such as 50, 100, or 300 dimensions. The vectors are initialized with small random numbers. **The embedding layer is used on the front end of a neural network and is fit in a supervised way using the Backpropagation algorithm.**²

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer

Count Vectorizer

Embedding

Classification
Algorithms

Naive Bayes

CNN

Linear SVM

Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning

Comparison

The End

²From the article What Are Word Embeddings for Text?

One-Hot Encoding for Word Embedding

1. Hello, how are you!
2. Win money, win from home.
3. Call me now
4. Hello, Call you tomorrow?

vocabulary = {are, call, from, hello, home, how, me, money, now, tomorrow, win, you}

Word	are	call	from	hello	home	how	me	money	now	tomorrow	win	you
Value	1	2	3	4	5	6	7	8	9	10	11	12
Sentence												
Hello, how are you!					4	6	1	12	0			
Win money, win from home.					11	8	11	3	5			
Call me now					2	7	9	0	0			
Hello, Call you tomorrow?					4	2	12	10	0			

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer

Count Vectorizer

Embedding

Classification
Algorithms

Naive Bayes

CNN

Linear SVM

Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning

Comparison

The End

Number of Parameters

Let's say that we want to embed sentences(or words) into a \mathbb{R}^n vector space. If m is the size of vocabulary, our Embedding layer has $m * n$ parameters that can be fitted in a supervised way using the Backpropagation.

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer

Count Vectorizer

Embedding

Classification
Algorithms

Naive Bayes

CNN

Linear SVM

Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning

Comparison

The End

Classification Algorithms

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

We used different classifiers and applied different models on the data. The classifiers we tested are:

- ▶ Naive Bayes
- ▶ Linear Support Vector Machine(SVM)
- ▶ Passive Aggressive Classifier
- ▶ Convolutional Neural Network(CNN)

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer
Count Vectorizer
Embedding

Classification
Algorithms

Naive Bayes
CNN
Linear SVM
Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning
Comparison

The End

Naive Bayes Classifier

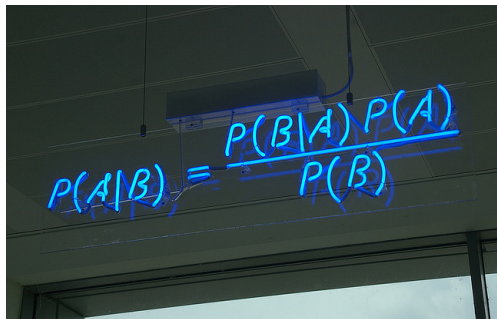

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Photo by Matt Buck



An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer

Count Vectorizer

Embedding

Classification
Algorithms

Naive Bayes

CNN

Linear SVM

Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning

Comparison

The End

Bayes Classifier: Naive One!

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

It is possible to show that accuracy is minimized, on average, by a very simple classifier that assigns each observation to the most likely class, given its predictor values. In other words, we should simply assign a test observation with predictor vector \mathbf{x}_0 to the class j for which

$$P(Y = j \mid \mathbf{X} = \mathbf{x})$$

is largest.

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

TF-idf Vectorizer

Count Vectorizer

Embedding

Classification
Algorithms

Naive Bayes

CNN

Linear SVM

Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning

Comparison

The End

Bayes Classifier: Naive One!

We make two assumptions:

1. X_1, X_2, \dots , and X_m are independent from each other;
2. $X_1, X_2, \dots, X_m \mid Y \sim MN(\cdot, p_1, p_2, \dots, p_m)$

$$\begin{aligned} P(Y=j \mid \mathbf{X} = (x_1, x_2, \dots, x_m)) &= \frac{P(\mathbf{X} = (x_1, x_2, \dots, x_m) \mid Y=j) \cdot P(Y=j)}{P(\mathbf{X} = \mathbf{x})} \\ &= \frac{P(X_1 = x_1 \mid Y=j) \cdot \dots \cdot P(X_m = x_m \mid Y=j) \cdot P(Y=j)}{P(\mathbf{X} = \mathbf{x})}. \end{aligned}$$

$$\begin{aligned} \hat{y} &= \arg \max_{j \in \text{classes}} \frac{P(X_1 = x_1 \mid Y=j) \cdot \dots \cdot P(X_m = x_m \mid Y=j) \cdot P(Y=j)}{P(\mathbf{X} = \mathbf{x})} \\ &= \arg \max_{j \in \text{classes}} P(X_1 = x_1 \mid Y=j) \cdot \dots \cdot P(X_m = x_m \mid Y=j) \cdot P(Y=j). \end{aligned}$$

Hyperparameters

Two important hyperparameters:

1. Size of the vocabulary;
2. Laplace/ Lidstone smoothing parameter(α).
3. Prior

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer
Count Vectorizer
Embedding

Classification
Algorithms

Naive Bayes
CNN
Linear SVM
Passive Aggressive Classifier

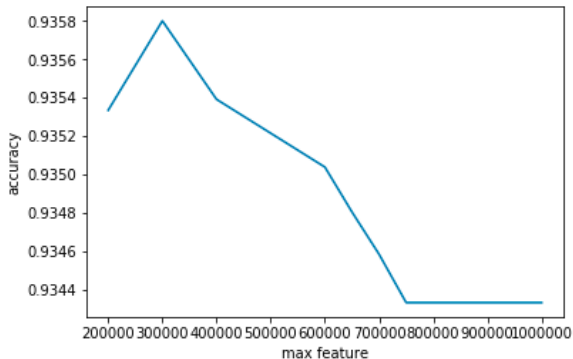
A Comparison
among Models

Ensemble Learning
Comparison

The End

Size of Vocabulary

We can determine the size of our vocabulary.



It is convex!

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer
Count Vectorizer
Embedding

Classification
Algorithms

Naive Bayes
CNN

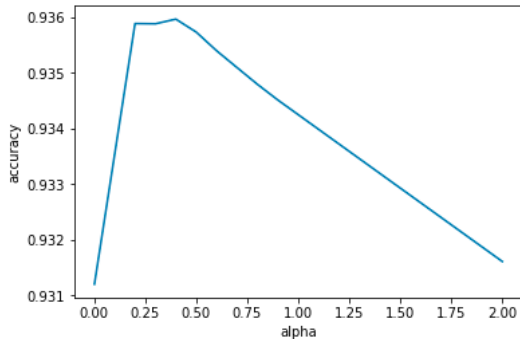
Linear SVM
Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning
Comparison

The End

Laplace/ Lidstone Smoothing Parameter(α)



It is convex!

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

- TF-idf Vectorizer
- Count Vectorizer
- Embedding

Classification
Algorithms

- Naive Bayes
- CNN
- Linear SVM
- Passive Aggressive Classifier

A Comparison
among Models

- Ensemble Learning
- Comparison

The End

Whether Use Prior or Not

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer

Count Vectorizer

Embedding

Classification
Algorithms

Naive Bayes

CNN

Linear SVM

Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning

Comparison

The End

$$P(c | x) = \frac{P(x | c) P(c)}{P(x)}$$

Posterior Probability

Likelihood

Class Prior Probability

Predictor Prior Probability

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \cdots \times P(x_n | c) \times P(c)$$

Whether Use Prior or Not

According to the dataset webpage, **distribution of dataset posts in different groups does not resemble the actual distributions**. So, if we fit a prior our accuracy with cross-validation increases, but it doesn't mean that our model is good; because model fits a wrong prior.

So we should not fit a prior(e.g. use an uninformative one).

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

TF-idf Vectorizer

Count Vectorizer

Embedding

Classification
Algorithms

Naive Bayes

CNN

Linear SVM

Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning

Comparison

The End

Bayes Classifier: Naive One!

Let's dive into code!

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer
Count Vectorizer
Embedding

Classification
Algorithms

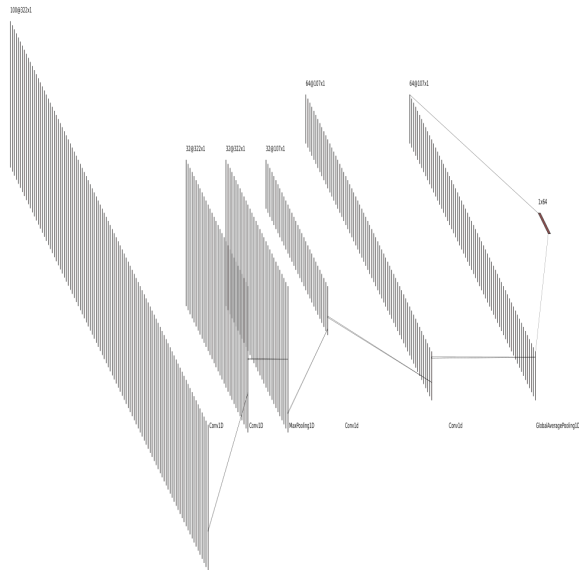
Naive Bayes
CNN
Linear SVM
Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning
Comparison

The End

CNN Over Embedding Layer



An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer
Count Vectorizer
Embedding

Classification
Algorithms

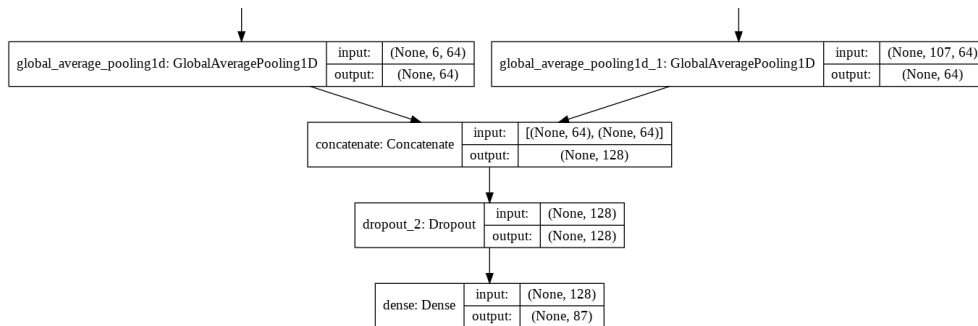
Naive Bayes
CNN
Linear SVM
Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning
Comparison

The End

CNN Over Embedding Layer



An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer
Count Vectorizer
Embedding

Classification
Algorithms

Naive Bayes
CNN
Linear SVM
Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning
Comparison

The End

CNN Over Embedding Layer

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer

Count Vectorizer

Embedding

Classification
Algorithms

Naive Bayes

CNN

Linear SVM

Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning

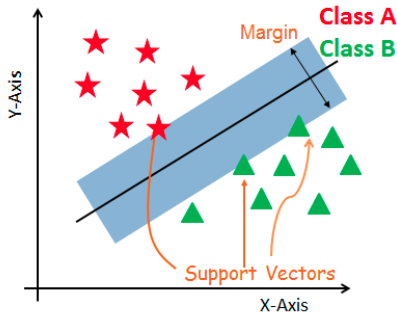
Comparison

The End

	سامسونگ	سونی	لوسٹر	پراید
سامسونگ	0	1.9844	6.6001	4.9251
سونی	1.9844	0	6.3962	4.8678
لوسٹر	6.6001	6.3962	0	5.8193
پراید	4.9251	4.8678	5.8193	0

Support Vector Machines

SVM constructs a hyperplane in multidimensional space to separate different classes. SVM generates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVM is to find a maximum marginal hyperplane(MMH) that best divides the dataset into classes.



An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

TF-idf Vectorizer
Count Vectorizer
Embedding

Classification
Algorithms

Naive Bayes
CNN

Linear SVM

Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning
Comparison

The End

How does SVM work?

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer

Count Vectorizer

Embedding

Classification
Algorithms

Naive Bayes

CNN

Linear SVM

Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning

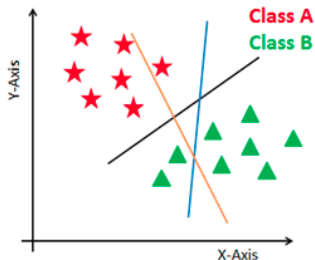
Comparison

The End

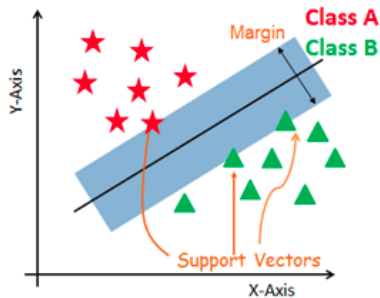
The main objective is to segregate the given dataset in the best possible way. The distance between the either nearest points is known as the margin. The objective is to select a hyperplane with the maximum possible margin between support vectors in the given dataset. SVM searches for the maximum marginal hyperplane in the following steps:

How does SVM work?

1. Generate hyperplanes which segregates the classes in the best way. Left-hand side figure showing three hyperplanes black, blue and orange. Here, the blue and orange have higher classification error, but the black is separating the two classes correctly.



2. Select the right hyperplane with the maximum segregation from the either nearest data points as shown in the right-hand side figure.



Linear SVM

Let's dive into code!

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer

Count Vectorizer

Embedding

Classification
Algorithms

Naive Bayes

CNN

Linear SVM

Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning

Comparison

The End

Passive Aggressive

- ▶ A margin based online learning algorithm
- ▶ Perfect for classifying massive streams
- ▶ Easy to implement and very fast
- ▶ **Passive**: if correct classification, keep the model;
- ▶ **Aggressive**: if incorrect classification, update to adjust to this misclassified example
- ▶ See <http://koaning.io/passive-aggressive-algorithms.html> for further reading

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer

Count Vectorizer

Embedding

Classification
Algorithms

Naive Bayes

CNN

Linear SVM

Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning

Comparison

The End

Passive Aggressive

Let's dive into code!

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer
Count Vectorizer
Embedding

Classification
Algorithms

Naive Bayes
CNN
Linear SVM

Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning
Comparison

The End

Ensemble Learning

- ▶ CountVectorizer + MultinomialNB
- ▶ TfidfVectorizer + MultinomialNB
- ▶ CountVectorizer + ComplementNB
- ▶ TfidfVectorizer + ComplementNB
- ▶ CountVectorizer + SVM(Hinge)
- ▶ CountVectorizer + SVM(HingeSq.)
- ▶ TfidfVectorizer + SVM(Hinge)
- ▶ TfidfVectorizer + SVM(HingeSq.)

PCA(100PC) +
Structured data

5-Layer Perceptron

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer
Count Vectorizer
Embedding

Classification
Algorithms

Naive Bayes
CNN
Linear SVM
Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning
Comparison

The End

Models Comparison

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

Tf-idf Vectorizer
Count Vectorizer
Embedding

Classification
Algorithms

Naive Bayes
CNN
Linear SVM
Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning
Comparison

The End

Name	Accuracy	Vectorizer	Classifier	Text Strategy
SVM	93.78	Tf-Idf	SVM	Dual Vectorizers
Ensemble	93.19	Count + Tf-Idf	Various!	Concat Text
P-A	92.80	Count	Passive-Agressive	Concat Text
CNN ³	90.50	Embedding	CNN	Dual CNN

³Trained with 80% of data!

Thanks for your attention!

- ▶ Codes and slides(in MLSP GitHub):
<https://github.com/ut-mlsp/Text-classification-crash-course>
- ▶ Divar posts dataset:
https://research.cafebazaar.ir/visage/divar_datasets/
- ▶ Any questions?

An Experience
with Text
Classification in
Datadays 2019

Majid Hajiheidari,
Amirmohammad
Asadi

Introduction:
Divar Dataset

The Problem:
Categorization

Feature Extraction

TF-idf Vectorizer
Count Vectorizer
Embedding

Classification
Algorithms

Naive Bayes
CNN
Linear SVM
Passive Aggressive Classifier

A Comparison
among Models

Ensemble Learning
Comparison

The End