Shahid Beheshti University

دانشگاه شهید بهشتی

دانشکده علوم ریاضی

گروه علوم کامپیوتر

گزارش تمرین های سری سوم

درس داده کاوی

اساتید محترم:

جناب آقای دکتر هادی فراهانی و جناب آقای دکتر سعید رضا خردپیشه

آموزشیار محترم: جناب آقای علی شریفی

جواد تدین **۹۹۴۲۲۰۴۱**

بهار ۱۴۰۰

به نام خدا

۱. در خصوص کرنل های پرکاربرد روش SVM تحقیق کنید .به صورت کلی چرا ما از ایده کرنل در بحث SVM بهره میبریم . آیا میتوان در خصوص کرنل ها و استفاده ی آنها حکم کلی داد . به طور مثال بگوییم از کرنل RBF در این مواقع خاص استفاده میکنیم .

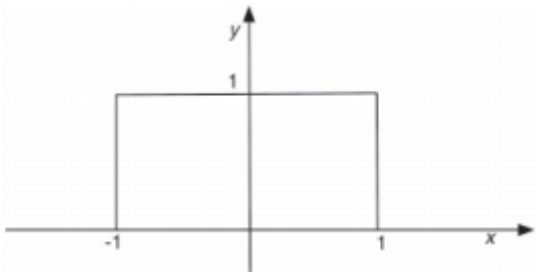
الگوریتم های SVM از مجموعه ای از توابع ریاضی که به عنوان کرنل تعریف می شوند، استفاده می کنند. وظیفه کرنل این است که داده ها را به عنوان ورودی گرفته و آن ها را به شکل مورد نیاز تبدیل کند .الگوریتم های مختلف SVM ، از انواع مختلف توابع کرنل ا ستفاده می کنند. این توابع می توانند انواع متفاوتی دا شته با شند. به عنوان مثال خطی ، غیرخطی ، چند جمله ای ، تابع پایه شعاعی(RBF) و سیگموئید.توابع کرنل ، برای داده های ترتیبی ،نمودار ها ، تصاویر و همچنین بردار ها معرفی می شوند .پرکاربردترین نوع تابع کرنل، RBF است. زیرا دارای پاسخ محلی و متناهی در کل بازه محور x است.توابع کرنل ، ضرب داخلی بین دو نقطه در یک فضای ویژگی مناسب را برمی گردانند. بنابراین ، با هزینه محاسباتی کم، حتی در فضاهای با ابعاد بالا، مفهومی از شباهت را تعریف می کنند.

قواعد کرنل

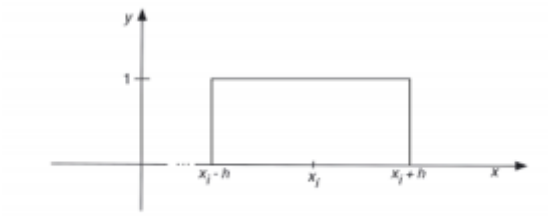
تعریف کرنل یا یک تابع پنجره به شرح زیر است:

$$K\left(\overline{x}\right)=\begin{matrix}1 & \text{if } \|\overline{x}\|\le 1\\ 0 & \text{otherwise}\end{matrix}$$

مقدار این تابع، در داخل یک شکل بسته به دامنه ۱ و مرکز مبدا مختصات برابر ۱ و در غیر این صورت ۰ است .همانطور که در شکل زیر نشان داده شده است:



برای xi ثابت، در داخل شکل بسته با دامنه h و مرکز xi ، تابع برابر ا ست با K(z-xi/h)=1و در غیر این صورت ۰ می با شد. همانطور که در شکل زیر نشان داده شده است:



بنابراین ، با انتخاب آرگومان K (.) پنجره را حرکت داده اید تا با دامنه h در مرکز xi قرار گیرد.

نمونه هایی از کرنل های SVM بیاید برخی از کرنل های رایج مورد استفاده در SVM ها و کاربرد های آن ها را مشاهده کنیم :

۱- کرنل چند جمله ای: این کرنل در پردازش تصویر پرکاربرد است. معادله آن به صورت زیر است:

$$k(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j} + 1)^d$$

که در آن d درجه چند جمله ای است .

۲- کرنل گاو سی: این یک کرنل برای اهداف عمومی ا ست. و هنگامی که هیچ دانش پی شینی در مورد داده ها وجود ندارد ا ستفاده می شود .معادله آن به صورت زیر است :

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

۳- تابع پایه شعاعی گاوسی (RBF): این کرنلی برای اهداف عمومی کلربرد دارد. و هنگامی که هیچ دانش پیشینی در مورد داده ها وجود نداشته باشد، مورد استفاده قرار می گیرد. معادله آن به صورت زیر است :

وبرای $\gamma > 0$ گاهی اوقات با استفاده از پاراماتر زیرمحاسبه می شود:

$$k(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma\|\mathbf{x_i} - \mathbf{x_j}\|^2)$$

۴- کرنلRBF لاپلاس: این هم یک کرنل برای اهداف عمومی است. و هنگامی که هیچ دانش پیشینی در مورد داده ها وجود ندارد استفاده می شود .معادله آن به صورت زیر است :

$$k(x, y) = \exp\left(-\frac{\|x - y\|}{\sigma}\right)$$

۵- کرنل تانژانت هیپربولیک (tanh) می توانیم از آن در شبکه های عصبی استفاده کنیم .معادله مربوط به آن عبارت است از:

$$k(\mathbf{x_i}, \mathbf{x_j}) = \tanh(\kappa\mathbf{x_i} \cdot \mathbf{x_j} + c)$$

در برخی موارد (نه همیشه $k > 0$ و $c < 0$)

۶- کرنل سیگموئید می توان این کرنل را در شبکه های عصبی مورد استفاده قرار داد. معادله مربوط به آن عبارت است از :

$$k(x, y) = \tanh(\alpha x^T y + c)$$

۷- کرنل تابع بسل (Bessel) از نوع اول ما می توانیم از آن برای حذف مقطع عرضی در توابع ریاضی استفاده کنیم. معادله آن عبارت است از :

$$k(x, y) = \frac{J_{v+1}(\sigma\|x - y\|)}{\|x - y\|^{-n(v+1)}}$$

۸- کرنل پایه شعاعی ANOVA ما می توانیم از آن در مسائل رگرسیون استفاده کنیم. معادله مربوط به آن عبارت است از:

$$k(x, y) = \sum_{k=1}^{n} \exp(-\sigma(x^k - y^k)^2)^d$$

۹- کرنلspline خطی بصورت یک بعدی این کرنل، هنگام کار با بردارهای بزرگ داده پراکنده ، کاربرد زیادی دارد. این کرنل اغلب در دسته بندی متن مورد استفاده قرار می گیرد. کرنل spline همچنین در مسائل رگرسیون عملکرد خوبی دارد. معادله آن عبارت است از:

$$k(x, y) = 1 + xy + xy\ min(x, y) - \frac{x + y}{2}\ min(x, y)^2 + \frac{1}{3}\min(x, y)^3$$

۲. قبلا با دیتاست کلاس بندی قیمت موبایل در کگل کار کرده ایم . بر روی دیتاست ، روش SVM را اجرا کنید . (استفاده از پکیج ها همانند sklearn مجاز است .)

jupyter Tadayon-project3-99422041 Last Checkpoint: 36 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

Run Code

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [148]:
```
data_set = pd.read_csv('mobile.csv')
data_set.head(10)
```

Out[148]:

	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	...	px_height	px_width	ram	sc_h	sc_w	talk_time	t
0	842	0	2.2	0	1	0	7	0.6	188	2	...	20	756	2549	9	7	19	
1	1021	1	0.5	1	0	1	53	0.7	136	3	...	905	1988	2631	17	3	7	
2	563	1	0.5	1	2	1	41	0.9	145	5	...	1263	1716	2603	11	2	9	
3	615	1	2.5	0	0	0	10	0.8	131	6	...	1216	1786	2769	16	8	11	
4	1821	1	1.2	0	13	1	44	0.6	141	2	...	1208	1212	1411	8	2	15	
5	1859	0	0.5	1	3	0	22	0.7	164	1	...	1004	1654	1067	17	1	10	
6	1821	0	1.7	0	4	1	10	0.8	139	8	...	381	1018	3220	13	8	18	
7	1954	0	0.5	1	0	0	24	0.8	187	4	...	512	1149	700	16	3	5	
8	1445	1	0.5	0	0	0	53	0.7	174	7	...	386	836	1099	17	1	20	
9	509	1	0.6	1	2	1	9	0.1	93	5	...	1137	1224	513	19	10	12	

10 rows × 21 columns

In [149]:
```
data_set.shape
```

Out[149]: (2000, 21)

In [150]:
```
X = data_set.drop('price_range', axis=1)
y = data_set['price_range']
```

In [151]:
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
```

In [152]:
```
from sklearn import svm
svm_clf = svm.SVC(kernel = 'linear')
svm_clf.fit(X_train, y_train)
```

Out[152]: SVC(kernel='linear')

In [153]:
```
from sklearn.metrics import accuracy_score, confusion_matrix
y_pred = svm_clf.predict(X_test)
cm = confusion_matrix(y_test,y_pred)
print('confusion matrix:\n',cm)
```

```
confusion matrix:
 [[145   0   0   0]
 [  4 152   1   0]
 [  0   5 152   1]
 [  0   0   4 136]]
```

In [154]:
```
sva =accuracy_score(y_test,y_pred)
print('accuracy score = ',accuracy_score(y_test,y_pred))
```

```
accuracy score =  0.975
confusion matrix:
 [[145   0   0   0]
 [  4 152   1   0]
 [  0   5 152   1]
 [  0   0   4 136]]
```

In [154]:
```
sva =accuracy_score(y_test,y_pred)
print('accuracy score = ',accuracy_score(y_test,y_pred))
```

```
accuracy score =  0.975
```

ابتدا دیتا ست را به دو قسمت test و train تقسیم می کنیم. و متغیر ها ی آموزش و برچسب y را ایجاد می کنیم و با پکیج sklearn روش svm را اجرا می کنیم .

accuracy score =0.975 به دست می آید یعنی داده های تست با ۰,۹۷ درصد برچسب ها را به درستی تشخیص داده است.

و با استفاده از ماتریس می توان دید که داده های کدام دسته به درستی تشخیص داده شده است.

۳. برای سوال ۲ حداقل ۵ حالت مختلف از قبیل کرنل ها و پارامترها را بررسی کنید و نتایج آن را گزارش دهید .

jupyter Tadayon-project3-99422041 Last Checkpoint: an hour ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

Run Code

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
```

In [152]:
```
from sklearn import svm
svm_clf = svm.SVC(kernel = 'linear')
svm_clf.fit(X_train, y_train)
```

Out[152]:
```
SVC(kernel='linear')
```

In [153]:
```
from sklearn.metrics import accuracy_score, confusion_matrix
y_pred = svm_clf.predict(X_test)
cm = confusion_matrix(y_test,y_pred)
print('confusion matrix:\n',cm)
```

```
confusion matrix:
 [[145   0   0   0]
 [  4 152   1   0]
 [  0   5 152   1]
 [  0   0   4 136]]
```

In [154]:
```
sva =accuracy_score(y_test,y_pred)
print('accuracy score = ',accuracy_score(y_test,y_pred))
```

```
accuracy score =  0.975
```

In [155]:
```
sva =accuracy_score(y_test,y_pred)
print('accuracy score = ',accuracy_score(y_test,y_pred))
```

```
accuracy score =  0.975
```

In [156]:
```
svm_clf = svm.SVC(kernel = 'poly', degree=3)
svm_clf.fit(X_train, y_train)
```

Out[156]:
```
SVC(kernel='poly')
```

In [157]:
```
y_pred = svm_clf.predict(X_test)
cm = confusion_matrix(y_test,y_pred)
print('confusion matrix:\n',cm)
```

```
confusion matrix:
 [[145   0   0   0]
 [  6 148   3   0]
 [  0   6 150   2]
 [  0   0   6 134]]
```

In [158]:
```
sva =accuracy_score(y_test,y_pred)
print('accuracy score = ',accuracy_score(y_test,y_pred))
```

```
accuracy score =  0.9616666666666667
```

In [159]:
```
svm_clf = svm.SVC(kernel = 'poly', degree=5)
svm_clf.fit(X_train, y_train)
```

Out[159]:
```
SVC(degree=5, kernel='poly')
```

۴. برای سوال ۲ سعی کنید مبحث soft margin و hard margin را بررسی کنید و نتایج آن را گزارش دهید .

زمانی که مقدار c را خیلی کم در نظر گرفتیم یعنی حاشیه ی ما خیلی نرم می شود و نسبت به خطاهای مدل حساسیت کمتری هم داریم پس خطاها بیشتر می شوند وoverfitting نیز رخ می دهد. ولی اگر زمانی c را مقدار خیلی بزرگتری در نظر بگیریم حاشیه سخت می شود و خطاها خیلی کمتر می شوند و ممکن است مرزی نتوانیم فیت کنیم همانطور که تغییرات c را می بینند

جواب در فایل کد

۵. مهندسی ویژگی یکی از بخش های مهم در فرایندهای علم داده میباشد . بر روی دیتاست موارد زیر را اجرا کنید .