降维：对手写数字集降维

1、导入所需模块和库

```
from sklearn.decomposition import PCA
from sklearn.ensemble import RandomForestClassifier as RFC
from sklearn.model_selection import cross_val_score
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

2、导入数据，探索数据

```
data = pd.read_csv('data/digit_recognizor.csv')
```

```
X = data.iloc[:,1:]
y = data.iloc[:,0]
```

```
X.shape
```

```
(42000, 784)
```

3、画累计方差贡献率曲线，找出最佳降维的维度的范围

```
%%time
pca = PCA()
pca_model = pca.fit(X)
```

```
Wall time: 4.87 s
```

```
## PCA的参数 n_components:
#1.为空 -- 实际的维数，没有降低维数
#2.整数 -- 降维到指定值 - 最大就是你的维度
#3.[0-1]之间的浮点数，另一个参数 svd_solver = 'full' ，，希望保留百分比的信息量
#4.'mle',最大似然估计自动选超参数的方法！！ -- 计算量相当可观
```

```
pca_model.explained_variance_ratio_ # 可解释性方差
```

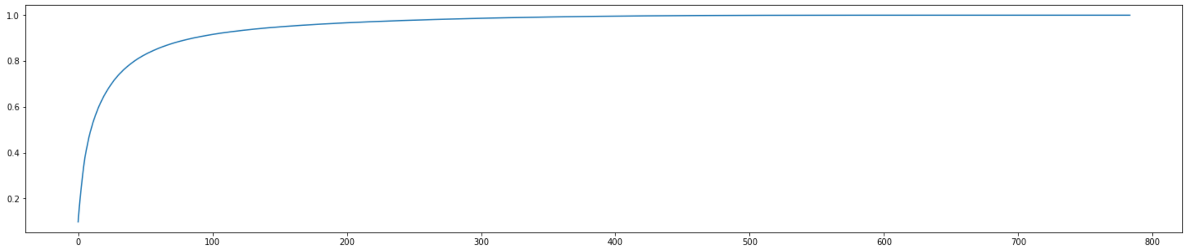
```
array([9.74893769e-02, 7.16026628e-02, 6.14590336e-02, 5.37930200e-02,
       4.89426213e-02, 4.30321399e-02, 3.27705076e-02, 2.89210317e-02,
       2.76690235e-02, 2.34887103e-02, 2.09932543e-02, 2.05900116e-02,
       1.70255350e-02, 1.69278702e-02, 1.58112641e-02, 1.48323962e-02,
       1.31968789e-02, 1.28272708e-02, 1.18797614e-02, 1.15275473e-02,
       1.07219122e-02, 1.01519930e-02, 9.64902259e-03, 9.12846068e-03,
       8.87640859e-03, 8.38766308e-03, 8.11855855e-03, 7.77405747e-03,
       7.40635116e-03, 6.86661489e-03, 6.57982211e-03, 6.38798611e-03,
       5.99367016e-03, 5.88913410e-03, 5.64335178e-03, 5.40967048e-03,
       5.09221943e-03, 4.87504936e-03, 4.75569422e-03, 4.66544724e-03,
       4.52952464e-03, 4.44989164e-03, 4.18255277e-03, 3.97505755e-03,
       3.84541993e-03, 3.74919479e-03, 3.61013219e-03, 3.48522166e-03,
       3.36487802e-03, 3.20738135e-03, 3.15467117e-03, 3.09145543e-03,
       2.93709181e-03, 2.86541339e-03, 2.80759437e-03, 2.69618435e-03,
       2.65831383e-03, 2.56298604e-03, 2.53821090e-03, 2.46178252e-03,
       2.39716188e-03, 2.38739578e-03, 2.27591447e-03, 2.21518444e-03,
       2.13933611e-03, 2.06133397e-03, 2.02851149e-03, 1.95976714e-03,
       1.93638614e-03, 1.88485334e-03, 1.86750994e-03, 1.81670044e-03,
       1.76891254e-03, 1.72592413e-03, 1.66120849e-03, 1.63309544e-03,
       1.60601340e-03, 1.54472396e-03, 1.46849742e-03, 1.42375935e-03,
       1.41098458e-03, 1.40228444e-03, 1.38834819e-03, 1.35417334e-03,
       1.32307196e-03, 1.30779914e-03, 1.29673803e-03, 1.24240433e-03,
       1.22249140e-03, 1.19624452e-03, 1.15840263e-03, 1.13858990e-03,
       1.12263468e-03, 1.10475151e-03, 1.08133084e-03, 1.07412608e-03,
       1.03865632e-03, 1.03321996e-03, 1.01494630e-03, 9.99965034e-04,
       9.74818997e-04, 9.45057663e-04, 9.38641819e-04, 9.12221022e-04,
       9.07313503e-04, 8.88871636e-04, 8.63699685e-04, 8.44234992e-04,
       8.35541599e-04, 8.16652064e-04, 7.87681237e-04, 7.81559589e-04,
       7.77464583e-04, 7.71933701e-04, 7.57842341e-04, 7.50215590e-04,
       7.34475822e-04, 7.25772099e-04, 7.15322651e-04, 7.00324985e-04,
       6.93049551e-04, 6.85741080e-04, 6.79933115e-04, 6.65718665e-04,
       6.56138536e-04, 6.44801786e-04, 6.35391815e-04, 6.26124921e-04,
       6.18514205e-04, 6.05741253e-04, 6.03854818e-04, 5.91452693e-04,
       5.85898805e-04, 5.84633227e-04, 5.75481878e-04, 5.69716770e-04,
       5.64497121e-04, 5.53174076e-04, 5.34344674e-04, 5.25777062e-04,
       5.21965550e-04, 5.11193757e-04, 5.05144243e-04, 4.99924977e-04,
       4.95323722e-04, 4.92348603e-04, 4.84395053e-04, 4.76686157e-04,
       4.74666022e-04, 4.67890879e-04, 4.65296674e-04, 4.61362258e-04,
       4.56336500e-04, 4.51756208e-04, 4.49488582e-04, 4.41357426e-04,
       4.38790136e-04, 4.24389337e-04, 4.20314615e-04, 4.16355388e-04,
       4.13830324e-04, 4.07099288e-04, 3.98548359e-04, 3.94359111e-04,
       3.94067067e-04, 3.91401338e-04, 3.81665552e-04, 3.79230943e-04,
       3.75536313e-04, 3.73831733e-04, 3.66337737e-04, 3.63074114e-04,
       3.59307646e-04, 3.57031988e-04, 3.53028613e-04, 3.52754592e-04,
       3.45302385e-04, 3.43276944e-04, 3.41890978e-04, 3.39165911e-04,
       3.34861511e-04, 3.29842492e-04, 3.25506437e-04, 3.24703554e-04,
       3.21013848e-04, 3.20432900e-04, 3.17194700e-04, 3.16819126e-04,
       3.10916530e-04, 3.10172960e-04, 3.06832953e-04, 3.03823171e-04,
       2.99708860e-04, 2.98327092e-04, 2.94093162e-04, 2.93391967e-04,
       2.92977220e-04, 2.89276406e-04, 2.84988966e-04, 2.83697252e-04,
       2.81082198e-04, 2.76258374e-04, 2.74103534e-04, 2.71602045e-04,
       2.67522376e-04, 2.66793948e-04, 2.62892231e-04, 2.62046588e-04,
       2.61487226e-04, 2.58253585e-04, 2.56603941e-04, 2.55387243e-04,
       2.54111610e-04, 2.52813861e-04, 2.50912463e-04, 2.48286117e-04,
       2.47615906e-04, 2.44461417e-04, 2.43286700e-04, 2.40992431e-04,
       2.40275940e-04, 2.39364452e-04, 2.38594369e-04, 2.36466045e-04,
       2.32057655e-04, 2.30987543e-04, 2.28982689e-04, 2.27286585e-04,
       2.25966858e-04, 2.24781908e-04, 2.20978300e-04, 2.19452970e-04,
       2.17383853e-04, 2.15927541e-04, 2.14951024e-04, 2.13614894e-04,
       2.11488233e-04, 2.10864069e-04, 2.08113405e-04, 2.05170657e-04,
       2.03746234e-04, 2.03271829e-04, 2.01784402e-04, 1.99666140e-04,
       1.97984405e-04, 1.96806855e-04, 1.94817073e-04, 1.93717714e-04,
       1.92666851e-04, 1.92124444e-04, 1.90249869e-04, 1.88388929e-04,
       1.86830011e-04, 1.84989673e-04, 1.84696295e-04, 1.84100838e-04,
       1.82788107e-04, 1.82323561e-04, 1.81392213e-04, 1.79612820e-04,
       1.76710787e-04, 1.75616351e-04, 1.74856503e-04, 1.73483732e-04,
       1.72524134e-04, 1.71476473e-04, 1.71133230e-04, 1.68773037e-04,
       1.68133359e-04, 1.67692755e-04, 1.66268327e-04, 1.64293501e-04,
       1.63902711e-04, 1.63026129e-04, 1.62247316e-04, 1.60434310e-04,
       1.60275519e-04, 1.58847778e-04, 1.58391347e-04, 1.57249091e-04,
       1.55421237e-04, 1.54449408e-04, 1.53626081e-04, 1.51833111e-04,
       1.50882402e-04, 1.50402146e-04, 1.49037531e-04, 1.48289379e-04,
       1.46698235e-04, 1.45918092e-04, 1.43470327e-04, 1.43413774e-04,
       1.42947318e-04, 1.41864032e-04, 1.41434477e-04, 1.40322317e-04,
       1.38342618e-04, 1.37446491e-04, 1.36378533e-04, 1.35450484e-04,
       1.35159928e-04, 1.34404079e-04, 1.33398331e-04, 1.32286985e-04,
       1.30752860e-04, 1.29890675e-04, 1.28535326e-04, 1.27848949e-04,
       1.26974506e-04, 1.26608300e-04, 1.25883934e-04, 1.25152668e-04,
       1.24026915e-04, 1.22541516e-04, 1.22323565e-04, 1.21184565e-04,
       1.20866127e-04, 1.20081813e-04, 1.19016518e-04, 1.18237887e-04,
       1.17234394e-04, 1.15741060e-04, 1.15360173e-04, 1.14560564e-04,
       1.13626694e-04, 1.13185761e-04, 1.11755777e-04, 1.10293511e-04,
       1.09960628e-04, 1.09349459e-04, 1.09176791e-04, 1.08477414e-04,
       1.07797530e-04, 1.06917463e-04, 1.06723498e-04, 1.06078717e-04,
       1.04464449e-04, 1.04396182e-04, 1.03069185e-04, 1.02214865e-04,
       1.01370888e-04, 1.00594605e-04, 9.97668874e-05, 9.94406703e-05,
       9.85829071e-05, 9.70159912e-05, 9.69784480e-05, 9.62936132e-05,
       9.46109643e-05, 9.41285257e-05, 9.30440801e-05, 9.25721602e-05,
       9.15809295e-05, 9.10792854e-05, 9.05018130e-05, 9.00300633e-05,
       8.96154897e-05, 8.89815850e-05, 8.83200096e-05, 8.75293892e-05,
       8.71255281e-05, 8.58975989e-05, 8.50483843e-05, 8.50402016e-05,
       8.46365570e-05, 8.38305712e-05, 8.27317681e-05, 8.25052090e-05,
       8.21602774e-05, 8.08585475e-05, 7.97090483e-05, 7.90024624e-05,
       7.83814628e-05, 7.80376415e-05, 7.74542297e-05, 7.66036440e-05,
       7.57515173e-05, 7.49309265e-05, 7.42332732e-05, 7.31760201e-05,
       7.30081526e-05, 7.25448153e-05, 7.22446811e-05, 7.17543752e-05,
       7.12784494e-05, 6.95487883e-05, 6.91951187e-05, 6.85545873e-05,
       6.78617549e-05, 6.70218240e-05, 6.64953450e-05, 6.58433865e-05,
       6.50054685e-05, 6.38837828e-05, 6.29081344e-05, 6.26693521e-05,
       6.13135178e-05, 6.07608686e-05, 6.02728272e-05, 5.94850515e-05,
       5.87997056e-05, 5.82619606e-05, 5.77536736e-05, 5.75026597e-05,
       5.62728232e-05, 5.60929317e-05, 5.49943468e-05, 5.44626534e-05,
       5.39581719e-05, 5.38260803e-05, 5.31310281e-05, 5.27592102e-05,
       5.21973100e-05, 5.13744317e-05, 5.12553126e-05, 5.06478383e-05,
       5.00844177e-05, 4.93501955e-05, 4.92647947e-05, 4.82612708e-05,
       4.72140977e-05, 4.68473649e-05, 4.61151761e-05, 4.59442191e-05,
       4.55818624e-05, 4.43711934e-05, 4.36168913e-05, 4.29144876e-05,
       4.26011409e-05, 4.24296094e-05, 4.18206054e-05, 4.11144641e-05,
       4.04988438e-05, 4.04029594e-05, 4.01015274e-05, 3.83703745e-05,
       3.79170304e-05, 3.76753121e-05, 3.74813570e-05, 3.72080430e-05,
       3.64897111e-05, 3.60866387e-05, 3.53090549e-05, 3.51349536e-05,
       3.47655497e-05, 3.44335233e-05, 3.40203237e-05, 3.37243003e-05,
       3.32093564e-05, 3.31208467e-05, 3.21412705e-05, 3.15566243e-05,
       3.12964929e-05, 3.08217792e-05, 3.01572208e-05, 2.98200940e-05,
       2.94476550e-05, 2.90894814e-05, 2.88892248e-05, 2.84426321e-05,
       2.82594076e-05, 2.78967823e-05, 2.68922180e-05, 2.66827861e-05,
       2.56238145e-05, 2.55681319e-05, 2.54091853e-05, 2.48281586e-05,
       2.47355902e-05, 2.43807400e-05, 2.42382680e-05, 2.40330977e-05,
       2.36180668e-05, 2.30583790e-05, 2.26794263e-05, 2.23684088e-05,
       2.17724416e-05, 2.13798456e-05, 2.12153679e-05, 2.11152854e-05,
       2.07851172e-05, 2.06477626e-05, 2.03117057e-05, 1.99614129e-05,
       1.97835385e-05, 1.94761131e-05, 1.89066352e-05, 1.88458128e-05,
       1.85580652e-05, 1.81787242e-05, 1.78865726e-05, 1.77108085e-05,
       1.74304017e-05, 1.72514805e-05, 1.61139955e-05, 1.58081384e-05,
       1.56945648e-05, 1.56593373e-05, 1.52931470e-05, 1.51817393e-05,
       1.50089148e-05, 1.47441220e-05, 1.44596323e-05, 1.43003593e-05,
       1.42052714e-05, 1.38929736e-05, 1.38569022e-05, 1.36791660e-05,
       1.34103769e-05, 1.32478815e-05, 1.31665557e-05, 1.25216577e-05,
       1.23306897e-05, 1.23143056e-05, 1.22392185e-05, 1.18470995e-05,
       1.16582089e-05, 1.16300724e-05, 1.15909253e-05, 1.14678450e-05,
       1.10729133e-05, 1.09294401e-05, 1.06581560e-05, 1.04485923e-05,
       1.03734264e-05, 1.03039169e-05, 9.85828293e-06, 9.49977979e-06,
       9.29870702e-06, 9.18567847e-06, 9.06529763e-06, 8.97446419e-06,
       8.91525355e-06, 8.57253818e-06, 8.51033741e-06, 8.18911115e-06,
       8.00253458e-06, 7.90484671e-06, 7.71865758e-06, 7.67024080e-06,
       7.44964443e-06, 7.39801923e-06, 7.27733058e-06, 7.05396957e-06,
       6.93978440e-06, 6.68594350e-06, 6.62761346e-06, 6.47609424e-06,
       6.44052795e-06, 6.08193709e-06, 6.02389863e-06, 5.83558228e-06,
       5.66293660e-06, 5.53890794e-06, 5.43930825e-06, 5.33622499e-06,
       5.20355108e-06, 5.17964590e-06, 5.10443954e-06, 5.00824157e-06,
       4.92763488e-06, 4.80114108e-06, 4.75510301e-06, 4.53930170e-06,
       4.35932679e-06, 4.28365458e-06, 4.21179602e-06, 4.04870947e-06,
       3.99717942e-06, 3.97306836e-06, 3.83255918e-06, 3.81301603e-06,
       3.77969020e-06, 3.55692331e-06, 3.40942146e-06, 3.37260841e-06,
       3.26669849e-06, 3.12204702e-06, 3.04952023e-06, 3.03881292e-06,
       2.98708906e-06, 2.83049255e-06, 2.70911629e-06, 2.66480551e-06,
       2.61840024e-06, 2.60619222e-06, 2.54442560e-06, 2.51639178e-06,
       2.44946063e-06, 2.36887034e-06, 2.31100440e-06, 2.26502791e-06,
       2.23470204e-06, 2.19711710e-06, 2.13734280e-06, 2.05838450e-06,
       1.94994618e-06, 1.87678350e-06, 1.78273495e-06, 1.76236299e-06,
       1.75750490e-06, 1.65331239e-06, 1.60778556e-06, 1.59117332e-06,
       1.57452615e-06, 1.48016712e-06, 1.43458910e-06, 1.39282403e-06,
       1.38289913e-06, 1.28228842e-06, 1.27857777e-06, 1.23930942e-06,
       1.20641201e-06, 1.18137630e-06, 1.13585335e-06, 1.08274212e-06,
       1.06559794e-06, 1.00410298e-06, 9.46188614e-07, 9.21884750e-07,
       8.72395784e-07, 8.61663770e-07, 8.36293489e-07, 8.27512338e-07,
       7.99545480e-07, 7.82822304e-07, 7.65897330e-07, 7.18126792e-07,
       6.98537214e-07, 6.95459660e-07, 6.91682829e-07, 6.53324249e-07,
       6.38094022e-07, 6.20802128e-07, 6.00107597e-07, 5.54533139e-07,
       5.44160769e-07, 5.19432084e-07, 5.11079815e-07, 4.95744841e-07,
       4.90011066e-07, 4.81912032e-07, 3.79703221e-07, 3.72190443e-07,
       3.62763214e-07, 3.52954381e-07, 3.23837840e-07, 3.23511964e-07,
       3.17446979e-07, 3.06885792e-07, 3.00230515e-07, 2.81547407e-07,
       2.48534247e-07, 2.45784501e-07, 2.33911869e-07, 2.30243100e-07,
       2.19018717e-07, 2.05624555e-07, 1.97663263e-07, 1.89810489e-07,
       1.89405730e-07, 1.86077809e-07, 1.75320734e-07, 1.70914908e-07,
       1.65198969e-07, 1.12310192e-07, 1.09654469e-07, 9.06587795e-08,
       8.72785660e-08, 7.50262391e-08, 7.48048733e-08, 6.89133157e-08,
       6.81449434e-08, 5.14838724e-08, 4.80326922e-08, 4.60969153e-08,
       4.27106616e-08, 3.94060178e-08, 3.80199097e-08, 3.23771252e-08,
       2.35171960e-08, 2.21403398e-08, 2.06834620e-08, 1.97134213e-08,
       1.93395216e-08, 1.76272370e-08, 1.70960746e-08, 1.51962633e-08,
       1.31700356e-08, 1.18156375e-08, 9.80662761e-09, 7.42553028e-09,
       7.07545401e-09, 5.75569441e-09, 3.81252053e-09, 3.30953194e-09,
       2.01269723e-09, 1.85705367e-09, 8.67397773e-10, 1.62308331e-10,
       1.06413966e-10, 8.86919537e-11, 1.75779111e-11, 1.60149185e-11,
       1.92313110e-32, 3.05514031e-33, 2.40291195e-33, 1.97721766e-33,
       1.70913353e-33, 1.61937649e-33, 1.36924730e-33, 8.74128832e-34,
       7.02332322e-34, 5.59777184e-34, 4.87954339e-34, 4.01309170e-34,
       4.01309170e-34, 4.01309170e-34, 4.01309170e-34, 4.01309170e-34,
       4.01309170e-34, 4.01309170e-34, 4.01309170e-34, 4.01309170e-34,
       4.01309170e-34, 4.01309170e-34, 4.01309170e-34, 4.01309170e-34,
       4.01309170e-34, 4.01309170e-34, 4.01309170e-34, 4.01309170e-34,
       4.01309170e-34, 4.01309170e-34, 4.01309170e-34, 4.01309170e-34,
       4.01309170e-34, 4.01309170e-34, 4.01309170e-34, 4.01309170e-34,
       4.01309170e-34, 4.01309170e-34, 4.01309170e-34, 4.01309170e-34,
       4.01309170e-34, 4.01309170e-34, 4.01309170e-34, 4.01309170e-34,
       4.01309170e-34, 4.01309170e-34, 4.01309170e-34, 4.01309170e-34,
       4.01309170e-34, 4.01309170e-34, 4.01309170e-34, 4.01309170e-34,
       4.01309170e-34, 4.01309170e-34, 4.01309170e-34, 4.01309170e-34,
       4.01309170e-34, 4.01309170e-34, 4.01309170e-34, 4.01309170e-34,
       4.01309170e-34, 4.01309170e-34, 4.01309170e-34, 4.01309170e-34,
       4.01309170e-34, 4.01309170e-34, 4.01309170e-34, 4.01309170e-34,
       4.01309170e-34, 4.01309170e-34, 4.01309170e-34, 4.01309170e-34,
       4.01309170e-34, 4.01309170e-34, 4.01309170e-34, 3.98635846e-34,
       2.59836120e-34, 8.99428984e-35, 7.17312032e-35, 1.82665870e-35])
```

```
pca_model.explained_variance_ratio_.sum()
```

```
1.0
```

```
plt.figure(figsize=[25,5])
plt.plot(np.cumsum(pca_model.explained_variance_ratio_))
```

```
[<matplotlib.lines.Line2D at 0x2890d4ee8e0>]
```



4.维度的学习曲线，搜索最佳维度的范围

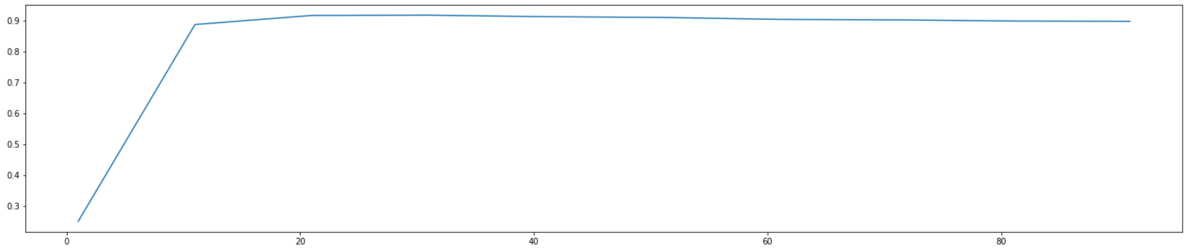
```
%%time

scores = []
for i in range(1,101,10):
    X_dr = PCA(i).fit_transform(X)
    score = 
cross_val_score(RFC(n_estimators=10,random_state=100),X_dr,y,cv=5).mean()
    scores.append(score)
```

```
Wall time: 2min 41s
```

```
plt.figure(figsize=[25,5])
plt.plot(range(1,101,10),scores)
```

```
[<matplotlib.lines.Line2D at 0x2890d59ed00>]
```



5、细化学习曲线

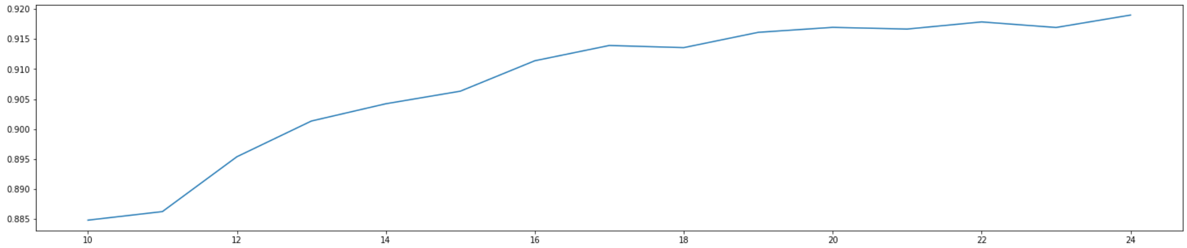
```
%%time

scores = []
for i in range(10,25):
    X_dr = PCA(i).fit_transform(X)
    score = 
cross_val_score(RFC(n_estimators=10,random_state=100),X_dr,y,cv=5).mean()
    scores.append(score)
```

```
Wall time: 2min 49s
```

```
plt.figure(figsize=[25,5])
plt.plot(range(10,25),scores)
```

```
[<matplotlib.lines.Line2D at 0x2890f26a2b0>]
```



6、在最佳维度下，查看模型效果

```
X_dr = PCA(24).fit_transform(X)
cross_val_score(RFC(n_estimators=10,random_state=100),X_dr,y,cv=5).mean()
```

```
0.9172857142857144
```

7、 尝试换一下模型

```
%%time

from sklearn.neighbors import KNeighborsClassifier as KNN
cross_val_score(KNN(),X_dr,y,cv=5).mean()
```

```
Wall time: 36.3 s
```

```
0.9703333333333333
```

8 KNN的k值的学习曲线

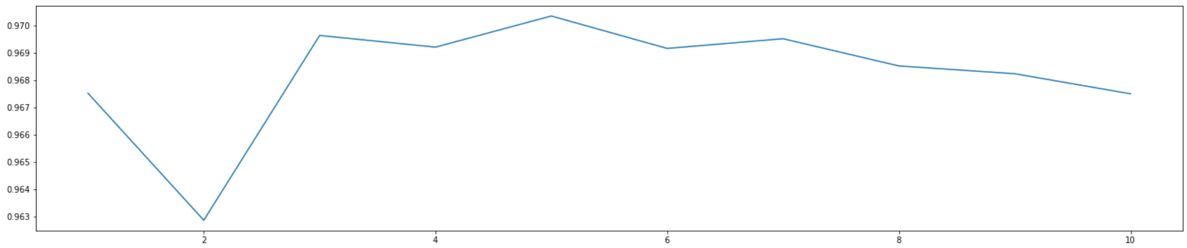
```
%%time

scores = []
for i in range(10):
    X_dr = PCA(24).fit_transform(X)
    score = cross_val_score(KNN(n_neighbors=i+1),X_dr,y,cv=5).mean()
    scores.append(score)
```

```
Wall time: 5min 49s
```

```
plt.figure(figsize=[25,5])
plt.plot(range(1,11),scores)
```

```
[<matplotlib.lines.Line2D at 0x2890f284190>]
```



```
cross_val_score(KNN(n_neighbors=5),X_dr,y,cv=5).mean()
#cross_val_score(KNN(5),X_dr,y,cv=5).mean()
```

```
0.9705238095238095
```

```
# 40 -> 1
```