Pattern Recognition

SPR_CHW2

K-Nearest Neighbors (KNN) For Classification

Mohammad Ali Basavad

40330475

بیان مسئله

در این تمرین قصد داریم با داشتن مجموعه ای از داده های اعداد(عکس اعداد یک رقمی). با استفاده از دسته بند KNN ، یک classifier ایجاد کنیم. در این تمرین قصد داریم با روش kfold ابتدا k مناسب برای دسته بند KNN را پیدا کرده سپس میزان دقت مدل براساس این classifier را حساب کنیم

دسته بند KNN

دسته بند KNN (k nearest neighbors) فاصله نمونه را با تمام داده های موجود محاسبه کرده و سپس به نمونه لیبل کلاسی را میزاند که بیشترین نزدیکی با دیتا های آن را داشته باشد.

در واقع در این روش K داده نزدیک به نمونه پیدا شده و نمونه به کلای تعلق پیدا می کند که بیشترین میزان را در K داده نزدیک به نمونه داشته باشد.

برای محاسبه فاصله نمونه تا داده ها می توان از روش های مختلقی ازجمله تابع cosine یا فاصله اقلیدسی استفاده کرد

$$consine = \cos(\theta) = \frac{A.B}{||A||\,.\,||B||}$$

$$Euclidean\ distance = d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

داده ها

در این تمرین از مجموعه داده openML mnist_784، استفاده خواهیم کرد. این دیتاست شامل عکس های 784 پیکسلی از رقم های مختلف می باشد که هر پیکسل باینری می باشد.

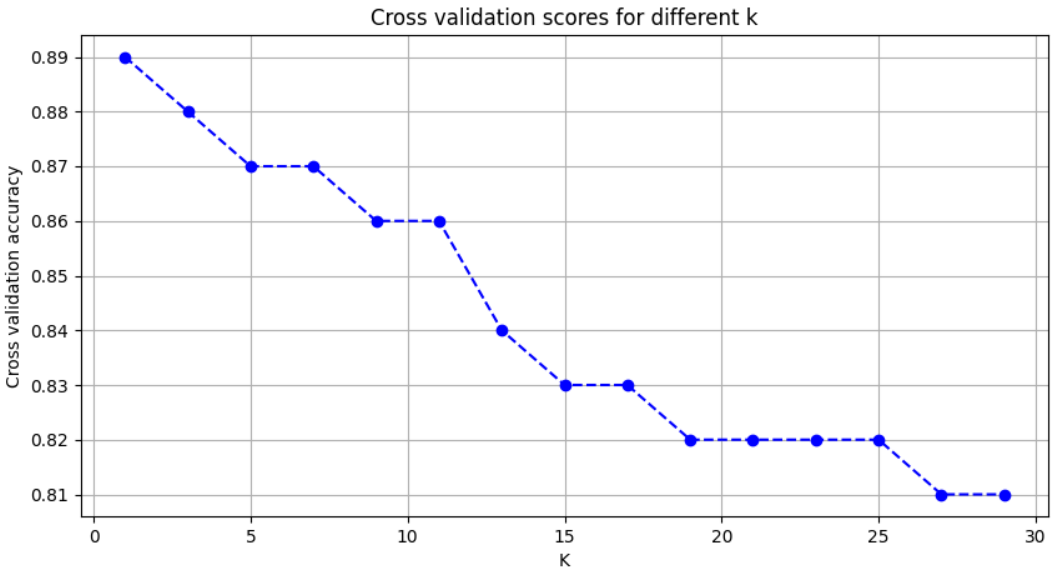
جداسازی داده ها

ابتدا 80٪ از داده های ابتدای دیتاست را بعنوان داده های Train و باقی 20٪ را بعنوان داده های Test جداسازی کردیم.

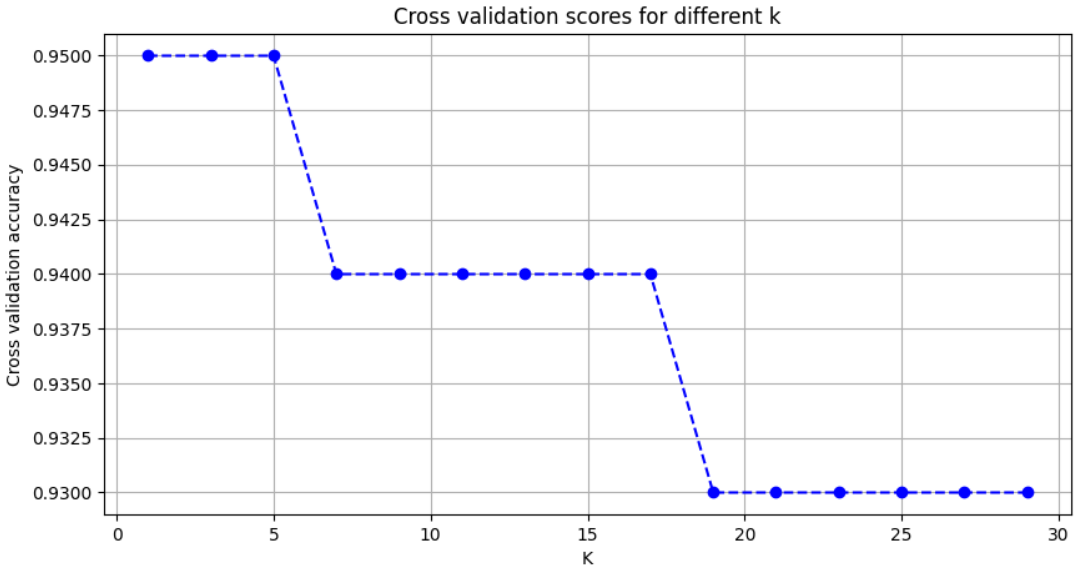
بدست آوردن K بهیینه

سپس برای پیدا کردن K بهیینه با استفاده از روش kfold داده های آموزش را به 10 قسمت تبدیل کرده هربار یک قسمت را بعنوان Test کنار گذاشته و با استفاده از باقی داده ها دسته بندی میکنیم و accuracy میانگین را برای 10 قسمت بدست آورده و اینکار را برای اعداد فرد بین 1 تا 29 بعنوان K ، تکرار میکنیم و برای هر K یک accuracy پیدا می شود، در نتیجه ما K بهیینه را از روی درصد دقت های بدست آمده پیدا می کنیم

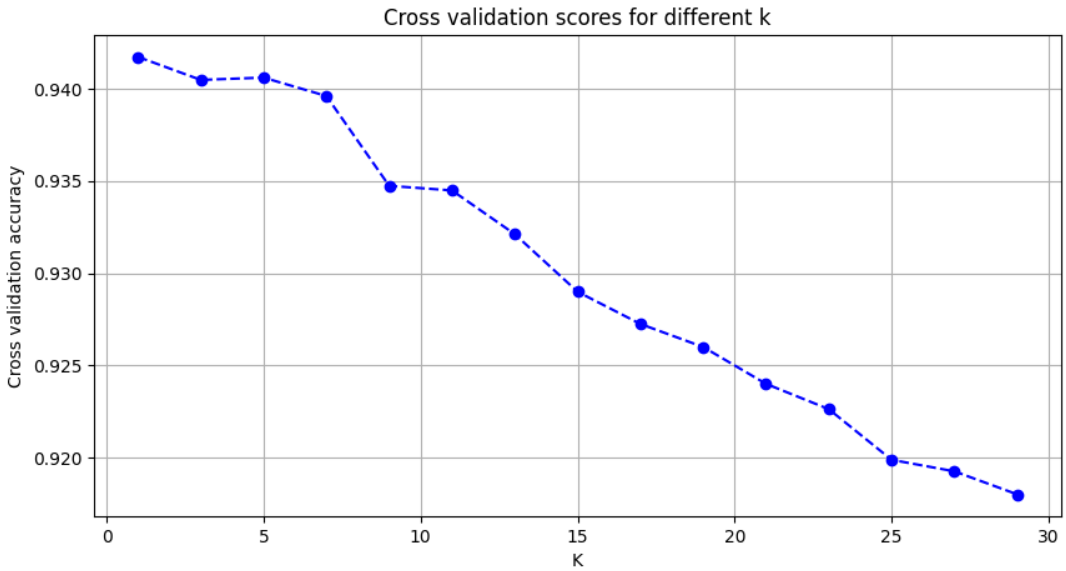
Cross validation using 1000 of datas and cosine metric



Cross validation using 10000 of datas and cosine metric



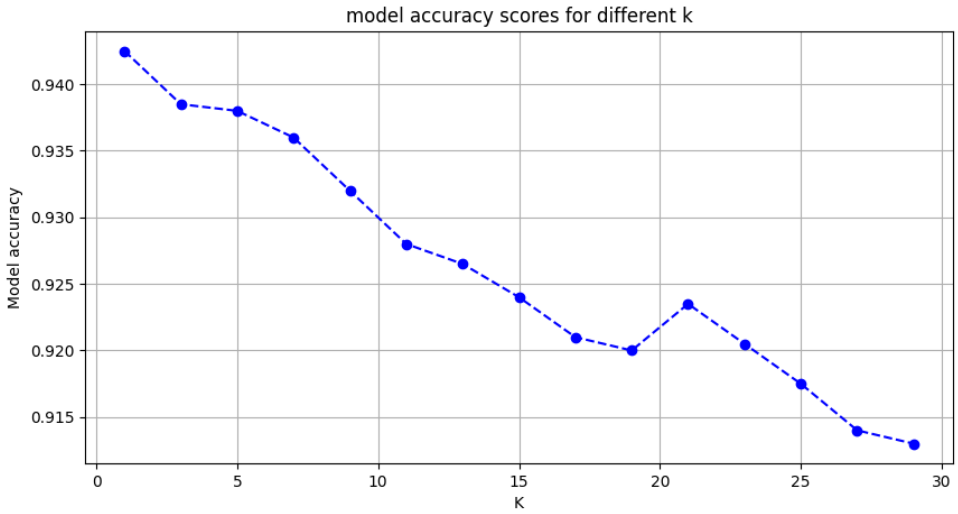
Cross validation using 10000 of datas and Euclidean metric



بدست آوردن دقت مدل

با توجه بدست آوردن K بهینه می توان از داده هایی که در ابتدا در دسته Test قرار دادیم برای بدست آوردن دقت نهایی مدل استفاده کرد

Model accuracy by K



نتیجه گیری

در نهایت می توان برای بهبود عملکرد زمانی مدل، میتوان با استفاده روش LeaveOneOut داده های Train را که به اشتباه دسته بندی می شوند را از مجموعه داده های خود حذف میکنیم. و بعد دوباره داده های Test خود را دسته بندی کنیم و میزان دقت مدل رو بسنجیم.

با استفاده از این روش زمان میانگین برای دسته بندی نمونه های جدید کاهش معنا داری پیدا می کند اما این روش مقداری از دقت الگوریتم کاهش میدهد

Accuracy in normal 1nn : `0.9425`

time in normal 1nn : `0.4150214195251465`

Accuracy in removeAmbiguousSamples 1nn : `0.939`

time in removeAmbiguousSamples 1nn : `0.33756589889526367`

After running each algorythm 50 times

Times in normal data :

[0.3538992404937744, 0.3870735168457031, 0.37909817695617676,
0.3584103584289551, 0.3645031452178955, 0.3682107925415039,
0.3520863056182861 3, 0.3610696792602539, 0.3524167537689209,
0.3589062690734863 3, 0.359588623046875, 0.3761718273162842,
0.3596780300140381, 0.3550672531127929 7, 0.3562066555023193 4,
0.349107027053833, 0.3656797409057617, 0.3680083751678467,
0.3544299602508545, 0.3521666526794433 6, 0.3600983619689941 4,
0.3586115837097168, 0.3549926280975342, 0.3555054664611816 4,
0.3579479484558105, 0.3558895587921142 6, 0.358096122741699 2,
0.3530433177947998, 0.3518576622009277 3, 0.3570451736450195 3,
0.3809812068939209, 0.3900940418243408, 0.3591651916503906,
0.3532023429870605 5, 0.3630421161651611 3, 0.3526856899261474 6,
0.3656044006347656, 0.352386951446533 2, 0.3602356910705566 4,
0.3524081707000732 4, 0.3583922386169433 6, 0.3723053932189941 4,
0.3569643497467041, 0.3557982444763183 6, 0.357041120529174 8,
0.3600430488586426, 0.3542890548706054 7, 0.355972290039062 5,
0.3665680885314941 4, 0.3922367095947265 6]

Time in removeAmbigousSample :

[0.3433394432067871, 0.3391196727752685 5, 0.3498015403747558 6,
0.3541240692138672, 0.341217517852783 2, 0.3474764823913574,
0.3972768783569336, 0.367123603820800 8, 0.354220151901245 1,
0.3452913761138916, 0.347916841506958, 0.347140550613403 3,
0.3605756759643554 7, 0.345265853259277 3, 0.351359605789184 57,
0.3511793613433838, 0.349442720413208, 0.335139513015747 07,
0.3512966632843017 6, 0.343740701675415 04, 0.341708466339111 3,
0.3500950336456299, 0.343339681625366 2, 0.341026782989501 95,
0.3465838432312011 7, 0.347902297973632 8, 0.340372323989868 16,
0.3432731628417969, 0.350013494491577 15, 0.344066381454467 8,
0.3547780513763428, 0.345463752746582 03, 0.342344284057617 2,
0.355391263961792, 0.342857122421264 65, 0.340317964553833,

0.35012340545654297, 0.34327220916748047, 0.34969258308410645,
0.34862852096557617, 0.3428654670715332, 0.3432917594909668,
0.347675085067749, 0.34937429428100586, 0.35387516021728516,
0.34657716751098633, 0.3445310592651367, 0.35310912132263184,
0.3477611541748047, 0.3728015422821045]

P_value = 1.64 * 10^-8

that shows there is meaningful difference between two methods

and second method is much better