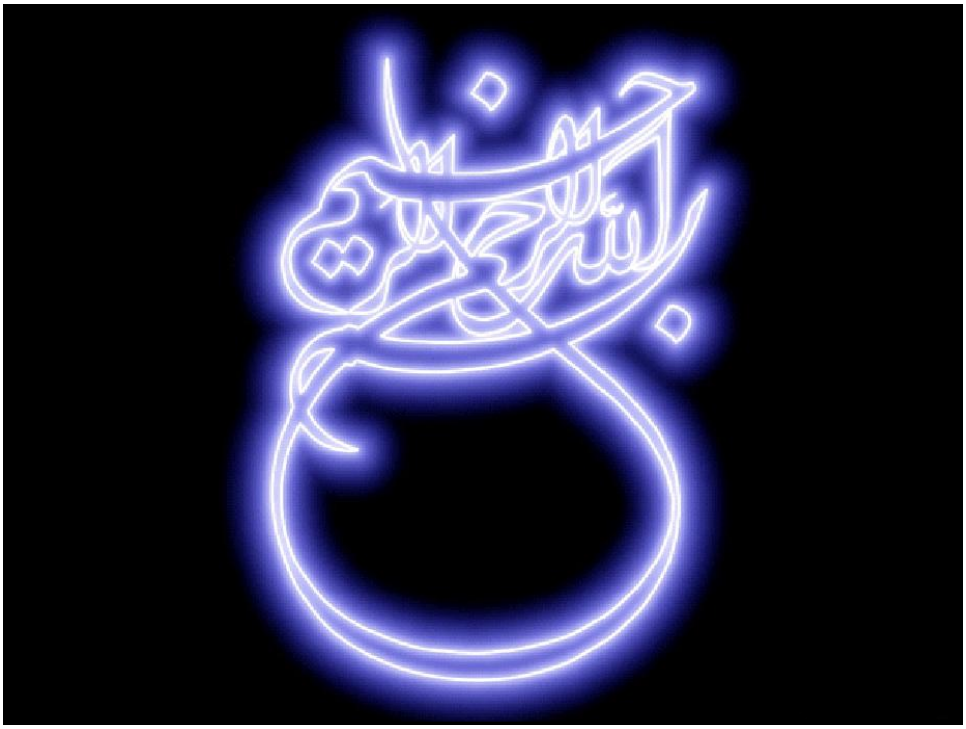


دسته بندی
بر مبنای نزدیکترین همسایه ها

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

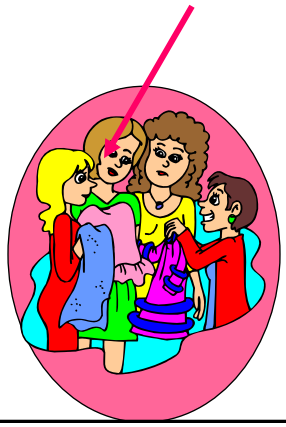
رئوس مطالب

- مفاهيم اوليه
- انواع classifier ها
- روال كلي KNN
- مشكلات و راه حلهاي ارائه شده

سميه عليزاده هيات علمي دانشكده صنايع
دانشگاه خواجه نصير طوسي

مفاهيم

Tell me who your friends are and I'll tell you who you are!



سميه عليزاده هيات علمي دانشكده صنايع
دانشگاه خواجه نصير طوسي

انواع دسته بندها

- Eager
  - ساخت مدل از نمونه هاي آموزشي قبل از classification
  - درخت تصميم گيري ، نمونه اي از eager classifier ها.
- Lazy
  - ذخيره سازي صرف نمونه هاي آموزشي ، بدون ساخت مدل از آنها
    - Instance based learner
    - تعويق يادگيري تا زمان classification
    - KNN نمونه اي از lazy classifier ها.
- تفاوت دو روش در زمان train و test

سميه عليزاده هيات علمي دانشكده صنايع
دانشگاه خواجه نصير طوسي

روال کلی KNN

1. تعيين پارامتر K (تعداد نزديكترين همسايه ها)
2. محاسبه **فاصله** نمونه ورودي با تمام نمونه هاي آموزشي
3. مرتب كردن نمونه هاي آموزشي بر اساس فاصله و انتخاب K همسايه نزديک
4. كلاسي كه **اكثريت** را در همسايه هاي نزديک دارد بعنوان تخميني براي كلاس نمونه ورودي بكار ببر (combine)

سميه عليزاده هيات علمي دانشكده صنايع
دانشگاه خواجه نصير طوسي

تابع فاصله

- براي محاسبه فاصله ميتوان از تابع فاصله اقليدسي استفاده كرد.
- فاصله اقليدسي بين دو تاپل X1 و X2 از رابطه زير بدست مي آيد.

$$\mathbf{X2}=(x_{21}; x_{22}; : : : ; x_{2n}) \text{ و } \mathbf{X1}=(x_{11}; x_{12}; : : : ; x_{1n})$$

$$dist(\boldsymbol{X_1}, \boldsymbol{X_2}) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}.$$

سميه عليزاده هيات علمی دانشكده صنايع
دانشگاه خواجه نصير طوسی

مثال

- در يک بررسي ، پرسشنامه اي براي دسته بندي كاغذ ها به دو دسته خوب و بد ، بر اساس دو ويژگي مقاومت در برابر اسيد و دوام انجام شد. 4 نمونه training در زير ديده ميشود:

X1 = مقاوت در برابر اسيد (seconds)	X2 = دوام (kg/square meter)	Y = Classification
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

كارخانه ، كاغذ جديدي توليد ميكند كه تست آزمايشگاه x1=3 و x2=7 را براي آن تعيين كرده است. ميخواهيم بدون تحقيق پرهزينه ، دسته بندي اين كاغذ را بدانيم.

سميه عليزاده هيات علمی دانشكده صنايع
دانشگاه خواجه نصير طوسی

مثال (ادامه...)

1. *تعیین پارامتر* K
فرض میکنیم $K = 3$

2. *محاسبه* ***فاصله*** *نمونه ورودي با تمام نمونه هاي آموزشي*
با در نظر گرفتن $(3,7)$ بعنوان ورودي فاصله آنرا با تمام نمونه هاي train محاسبه میکنیم.

مقاومت در برابر اسید = X1 (seconds)	دوام = X2 (kg/square meter)	**فاصله اقلیدسی با نمونه (3, 7)**
7	7	$\sqrt{(7-3)^2+(7-7)^2}=4$
7	4	$\sqrt{(7-3)^2+(4-7)^2}=5$
3	4	$\sqrt{(3-3)^2+(4-7)^2}=3$
1	4	$\sqrt{(1-3)^2+(4-7)^2}=\sqrt{13}$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مثال (ادامه...)

3. مرتب کردن نمونه هاي آموزشي بر اساس فاصله و انتخاب K همسایه نزدیک

مقاومت در برابر اسید = X1 (seconds)	دوام = X2 (kg/square meter)	فاصله اقلیدسی با نمونه (3, 7)	**رتبه (فاصله اقلیدسی)**	**جزو ۳ همسایه نزدیک هست؟**
7	7	$\sqrt{(7-3)^2+(7-7)^2}=4$	۳	بله
7	4	$\sqrt{(7-3)^2+(4-7)^2}=5$	٤	خیر
3	4	$\sqrt{(3-3)^2+(4-7)^2}=3$	۱	بله
1	4	$\sqrt{(1-3)^2+(4-7)^2}=\sqrt{13}$	۲	بله

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مثال (ادامه...)

4. کلاسي که **اکثريت** را در همسايه هاي نزديک دارد بعنوان تخميني براي کلاس نمونه ورودي بکار ببر.

کلاس نزدیکترین همسایه	جزو ۳ همسایه نزدیک هست ؟	رتبه (فاصله اقلیدسی)	فاصله اقلیدسی با نمونه (3, 7)	X2 = دوام (kg/square meter)	X1 = مقاومت در برابر اسید (seconds)
Bad	بله	۳	$\sqrt{(7-3)^2+(7-7)^2}=4$	7	7
-	خیر	٤	$\sqrt{(7-3)^2+(4-7)^2}=5$	4	7
Good	بله	۱	$\sqrt{(3-3)^2+(4-7)^2}=3$	4	3
Good	بله	۲	$\sqrt{(1-3)^2+(4-7)^2}=\sqrt{13}$	4	1

چون دو Good و یک Bad داریم و 1<2 نتیجه میگیریم که کاغذي که در آزمايشگاه اعداد x1=3و X2=7 را بدست آورده است در دسته خوب قرار ميگيرد.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

بررسی دقیق تر KNN

- مسائل مربوط به تابع فاصله
- انتخاب تابع combine
- انتخاب مقدار K

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مسائل مربوط به تابع فاصله

1. ويژگيهاي غير عددي
2. تفاوت در مقياس اندازه گيري
3. عدم وجود مقدار براي ويژگي
4. انتخاب تابع فاصله
5. ويژگيهاي نا مرتبط

سميه عليزاده هيات علمی دانشكده صنايع
دانشگاه خواجه نصير طوسی

مسائل مربوط به تابع فاصله

- *مقايسه مقادير ويژگيهاي غيرعددي*

  - ساده ترين روش مقايسه اين است كه اگر مقدار ويژگي در دو نمونه برابر است تفاوت 0 و در غير اينصورت تفاوت 1 در نظر گرفته شود ؟
  - وجود روشهاي پيچيده تر

سميه عليزاده هيات علمی دانشكده صنايع
دانشگاه خواجه نصير طوسی

مسائل مربوط به تابع فاصله (ادامه...)

- *تفاوت در مقیاس اندازه گیري ویژگیها*
  - *محو اثر بعضي ویژگیها*

- ابتدا همه مقادیر نرمال شده و سپس مقایسه صورت میگیرد
- نرمالسازي ویژگي A با مقدار v به مقدار v' در بازه [0,1]

$$v' = \frac{v - min_A}{max_A - min_A}$$

min_A و max_A مینیمم و ماکزیمم روي مجموعه train محاسبه میشود.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مسائل مربوط به تابع فاصله (ادامه...)

- *ویژگي در یک (یا دو) نمونه مقدار ندارد*
  - حداکثر مقدار ممکن بعنوان تفاوت نظر گرفته میشود.

ویژگی	تفاوت
غیرعددی	۱
عددی	
هیچکدام مقدار ندارد	۱
در یکي مقدار v دارد و دیگري مقدار ندارد	مقدار بزرگتر v و $1-v$ در نظر گرفته میشود.

  - مثال : ویژگي A در یک نمونه مقدار 0.4 و در دیگري مقدار ندارد
$$\max(0.4, 1-0.4) = 0.6$$
0.6 بعنوان تفاوت در نظر گرفته میشود

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مسائل مربوط به تابع فاصله (ادامه...)

- *انتخاب تابع فاصله*

فاصله بين دو نمونه $(x_{1n} ; : : : ; x_{12}; x_{11})$=**X1** و $(x_{2n} ; : : : ; x_{22}; x_{21})$=**X2** ميتواند از روابط زير محاسبه شود :

1. Manhattan

2. تابع اقليدسي

$$dist(\mathbf{x_1}, \mathbf{x_2}) = \sum_{i=1}^{n} |x_{1i} - x_{2i}|$$

$$dist(\mathbf{x_1}, \mathbf{x_2}) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}$$

- سخت گيري
- صرف نظر از ريشه
- استفاده از توانهاي ديگر
- تابع مرسوم

سميه عليزاده هيات علمي دانشكده صنايع
دانشگاه خواجه نصير طوسي

مسائل مربوط به تابع فاصله (ادامه...)

- *يكسان گرفتن اهميت ويژگيها در تابع فاصله*
  - وجود ويژگيهاي نامرتبط
- لزوم در نظر گرفتن وزن براي ويژگيها

$$Euclidean(X1, X2) = \sqrt{\sum_{i=1}^{n} w_i (x_{1i} - x_{2i})^2}$$

- روشهاي تعيين وزن
  - نظر خبره
  - الگوريتمي

سميه عليزاده هيات علمي دانشكده صنايع
دانشگاه خواجه نصير طوسي

مسائل مربوط به تابع فاصله (ادامه...)

- روش هاي تعيين وزن الگوريتمي
  - Cross-validation –
    - مقدار تصادفي اوليه براي وزن ويژگيها ، انجام دسته بندي و تغيير براي مينيمم كردن نرخ خطا
    - الگوريتم ژنتيک
  - روش Aha –
    - با یک مقدار اولیه شروع میکند و بعد از هر بار دسته بندي ، وزن ها را عوض میکند.
  - شرط درستي روش –
    - با اضافه كردن ويژگيهاي نامرتبط ، كيفيت دسته بندي افت نكند

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

روشهای ارزیابی دسته بندی ها

- صحت پيش بيني (Accuracy)
- زمان لازم براي ساخت مدل و زمان لازم براي استفاده از مدل
- پايداري (Robustness):
  توانايي برخورد با داده هاي غير معمول (noise) و يا مقادير جا افتاده (missing values)
- قابليت تفسير (Interpretability):
  قابل فهم بودن مدل و فراهم سازي بينش لازم توسط آن
- جمع و جور بودن (Compactness) مدل :
  اندازه درخت يا تعداد قواعد و ...

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

شاخص دقت در دسته بندی

Classes	*buys_computer = yes*	*buys_computer = no*	*Total*	*Recognition (%)*
buys_computer = yes	6,954	46	7,000	99.34
buys_computer = no	412	2,588	3,000	86.27
Total	7,366	2,634	10,000	95.52

	C_1	C_2
C_1	true positives	false negatives
C_2	false positives	true negatives

حساسیت

$$sensitivity = \frac{t\_pos}{pos}$$

شفافیت

$$specificity = \frac{t\_neg}{neg}$$

دقت

$$precision = \frac{t\_pos}{(t\_pos + f\_pos)}$$

صحت

$$accuracy = sensitivity \frac{pos}{(pos + neg)} + specificity \frac{neg}{(pos + neg)}.$$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی