

یادگیری ماشین
Machine Learning
مهر ۱۴۰۴

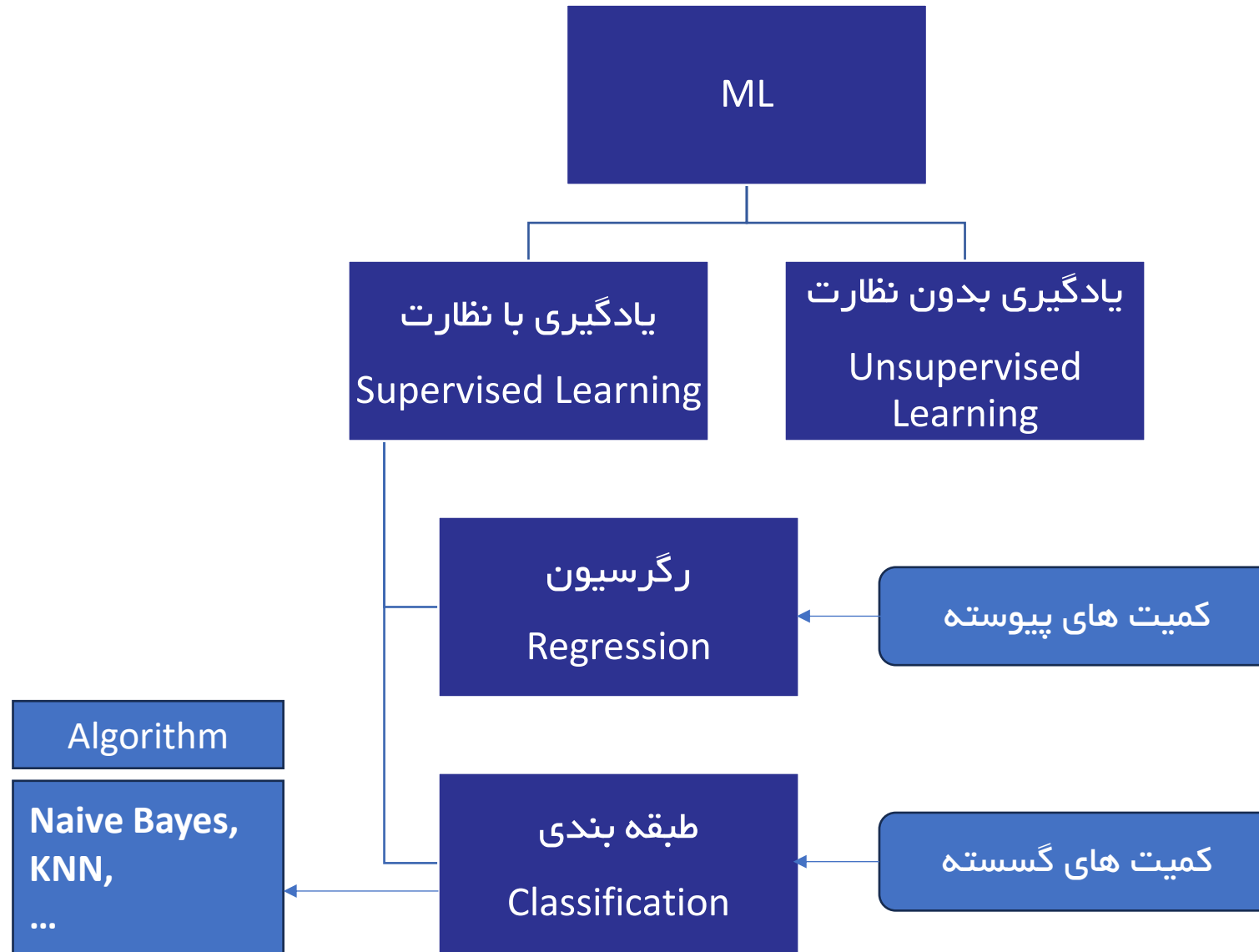
KNN Algorithm

k-nearest neighbors

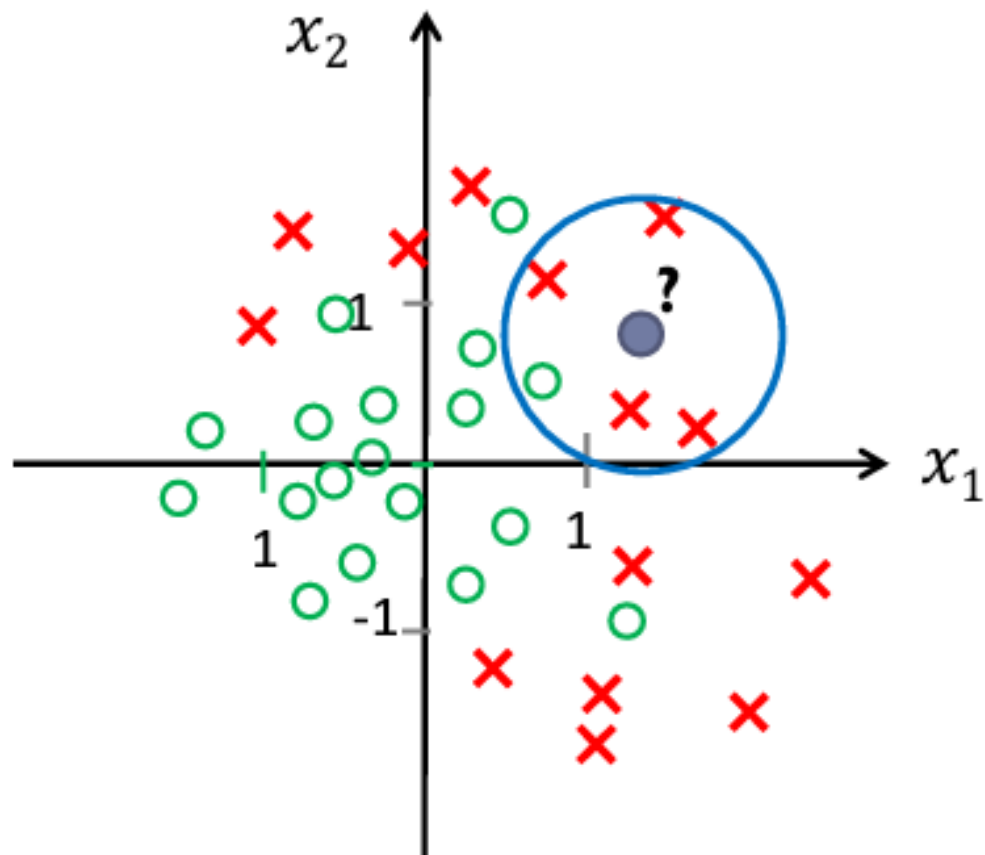
استاد : محسن عباسی



موسسه آموزش عالی خراسان



KNN



ایده ی اصلی

برای پیش‌بینی برچسب یک داده ی جدید، به k تا از

«نزدیک‌ترین» داده‌های آموزش نگاه می‌کنیم و بر اساس رأی

اکثریت (در طبقه‌بندی) خروجی را تعیین می‌کنیم.

KNN Algorithm

1

محاسبه ی فاصله

بین نقطه ی جدید و همه ی
نمونه های آموزش
(معمولاً فاصله اقلیدسی)



2

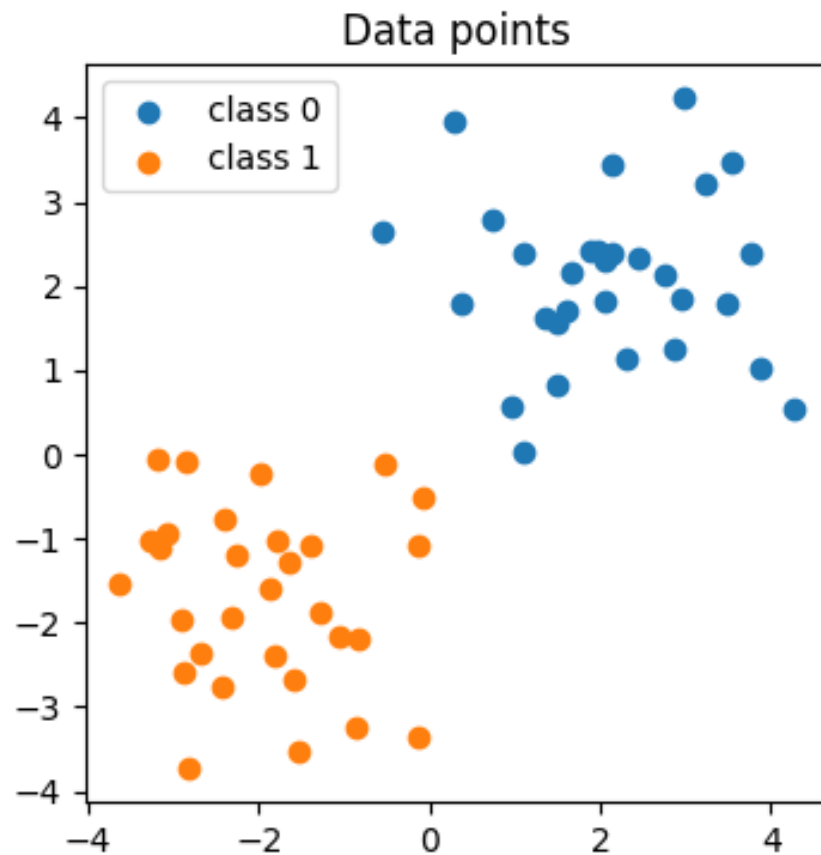
نمونه ی k انتخاب
نزدیکتر



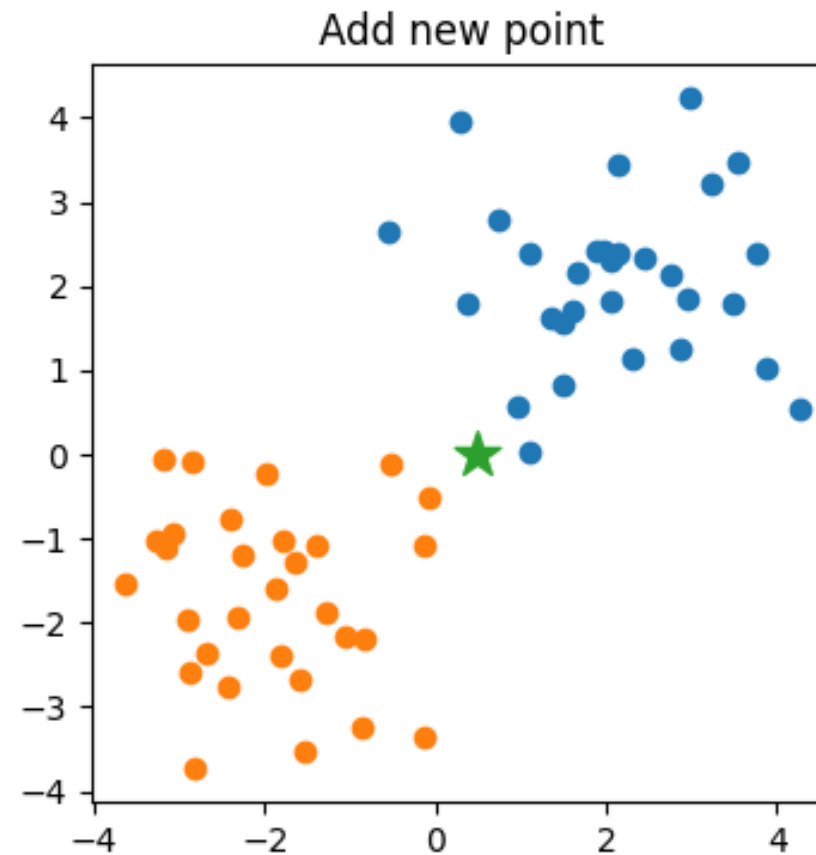
3

خروجی گیری

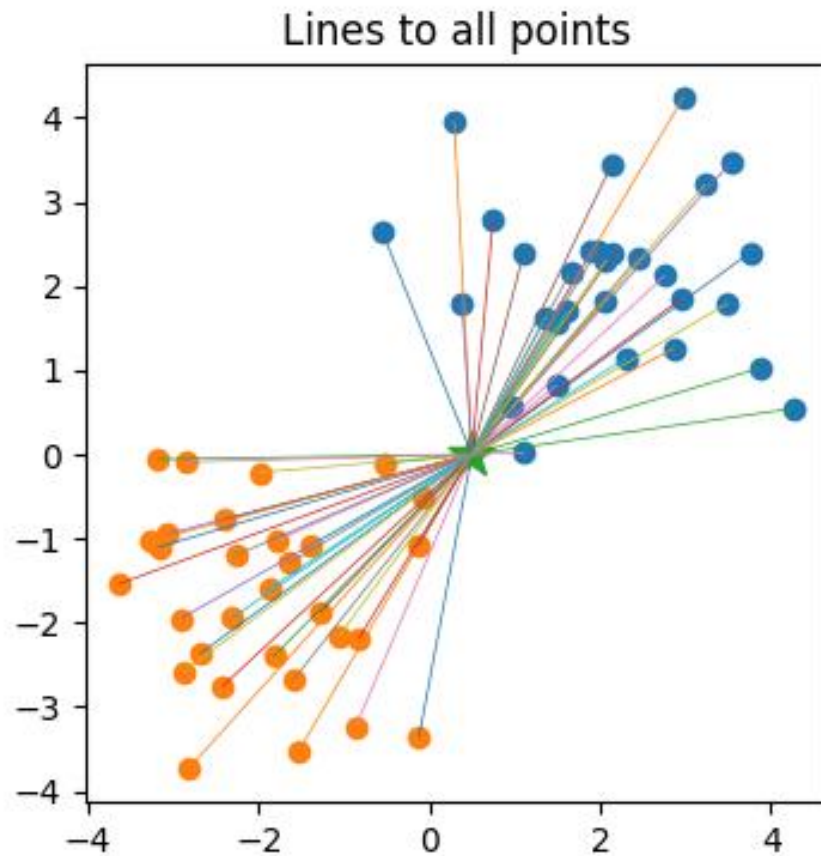
رای گیری بین برچسب ها



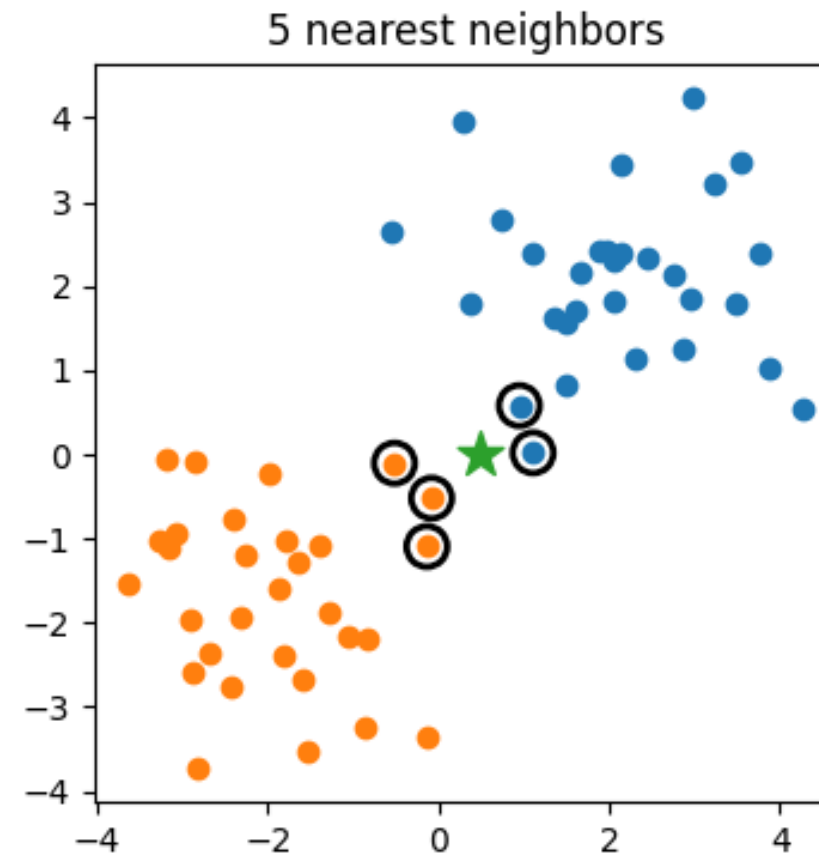
۱- دیتای آموزش
بر اساس برچسب



۲- ورود دیتای تست



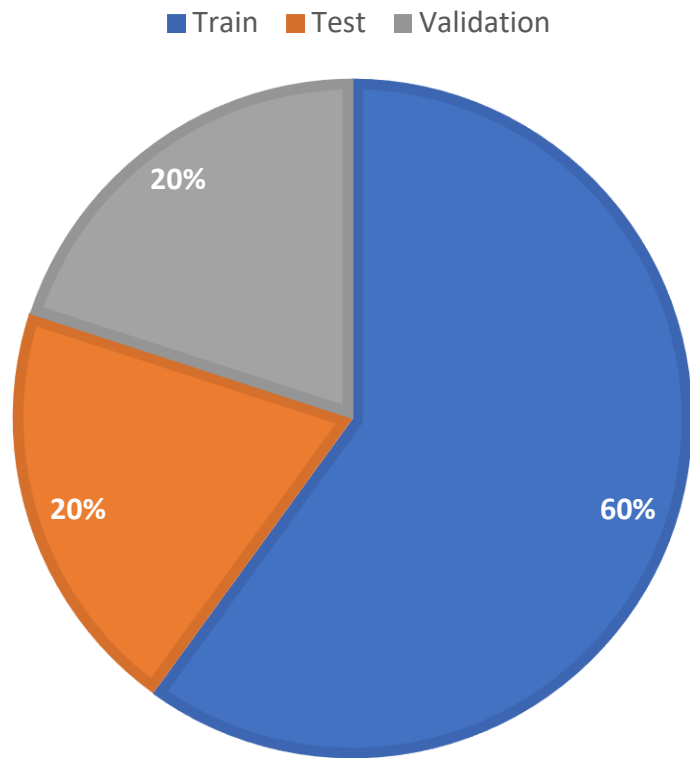
۳- محاسبه فاصله تمام نقاط آموزش
تا نقطه تست



۴- یافتن $k=5$ همسایه دیتای تست

الگوریتم تنبل (در مرحله آموزش اتفاقی نمی افتد و همه محاسبات در حین پیش بینی محاسبه میشود.)
کند روی دیتاست‌های بزرگ (نیاز به محاسبه فاصله با همه)
حساس به مقیاس ویژگی‌ها (نیاز به نرمال‌سازی)
کارایی ضعیف در ابعاد زیاد curse of dimensionality

Validation



تقسیم داده به Train / Validation / Test

برای ارزیابی منصفانه ی مدل و جلوگیری از overfitting، داده به سه بخش جدا تقسیم می شود:

1. Train | آموزش

مدل با استفاده از این بخش یاد می گیرد و پارامترهای داخلی اش تنظیم می شود.

2. Validation | اعتبارسنجی

برای انتخاب و تنظیم هایپرپارامترها (مثلاً مقدار k در KNN یا نرخ یادگیری)، بدون اینکه مدل داده ی تست را ببیند.

3. Test | آزمون نهایی

فقط یک بار در انتهای کار برای سنجش نهایی عملکرد مدل روی داده ی دیده نشده.

استفاده از تست در طول توسعه ممنوع است تا گزارش نهایی واقعی باشد.

انتخاب k

k کوچک \rightarrow مدل به نویز حساس می‌شود، احتمال $overfitting$ بالا
 k بزرگ \rightarrow تصمیم‌گیری نرم‌تر اما احتمال $underfitting$ و کاهش دقت روی جزئیات
معمولاً انتخاب k با روش‌های $Cross-Validation$ انجام می‌شود
مقدار k اغلب عددی فرد (برای جلوگیری از تساوی در رأی‌گیری) انتخاب می‌شود
انتخاب k وابسته به تراکم و اندازه ی داده است؛ با افزایش داده ی آموزشی، معمولاً k هم کمی بزرگ‌تر انتخاب می‌شود

F1 Score

F1-Score یک معیار ترکیبی برای ارزیابی مدل‌های طبقه‌بندی است.

تعدادل بین Precision و Recall را نشان می‌دهد:

$$F = \frac{1}{\frac{1}{2P} + \frac{1}{2R}} = \frac{2PR}{P + R}$$

Precision: درصد پیش‌بینی‌های درست مثبت نسبت به کل پیش‌بینی‌های مثبت

Recall: درصد پیش‌بینی‌های درست مثبت نسبت به کل نمونه‌های مثبت واقعی

F1-Score نزدیک به ۱ \leftarrow مدل دقیق و حساس است

F1-Score نزدیک به ۰ \leftarrow مدل عملکرد ضعیف دارد

۱- انتخاب دیتاست

هر نفر یک دیتاست مولتی کلاس انتخاب کند. (با حداقل ۳ کلاس) ویژگی‌ها و برچسب‌ها مشخص باشند.

۲- تقسیم‌بندی داده

داده به سه بخش تقسیم شود:
validation – train – test

۳- اجرای KNN

انتخاب k با استفاده از داده validation

۴- پیاده‌سازی مدل KNN

۵- ارزیابی مدل

محاسبه معیارهای Accuracy و F1-Score

۶- رسم Confusion Matrix

۷- گزارش نهایی

انتخاب k به چه صورت انجام شد
معیارهای عملکرد مدل

تحلیل ماتریس در هم‌ریختگی: چه کلاس‌هایی با هم اشتباه گرفته شدند؟

پروژه

GitHub



https://github.com/vahidseyyedi/ML_uni

تهیه کننده:
وحید سیدی