ورود به دنیای هوش‌مصنوعی

دوره‌ی مقدماتی و رایگان برای عموم علاقه‌مندان

جلسه‌ی ششم: ارزیابی دسته‌بندی، الگوریتم K نزدیک‌ترین همسایه

علی شریفی زارچی

رئیس کمیته‌ی علمی بین‌المللی المپیاد جهانی هوش‌مصنوعی

عضو کمیته‌ی علمی بین‌المللی المپیاد جهانی کامپیوتر

مرور جلسه‌ی قبل

- پیش‌بینی برچسب گسسته: دسته‌بندی (Classification)
- درخت تصمیم (Decision Tree)
- مجموعه داده‌ی کشتی تایتانیک
- عمق درخت: تعداد پارامترها
- بیش‌برازش و کم‌برازش درخت تصمیم
- بارگذاری و پیش‌پردازش داده‌ها در پایتون

تصحیح جلسه‌ی پنجم: یک درخت تصمیم روی تایتانیک

درصدهای بقا از داده‌ی واقعی تایتانیک؛ درخت با دو-سه سوال ساده به پیش‌بینی خوبی می‌رسد و کاملا قابل توضیح است.

سوال کلیدی

مدل ساختیم و پیش‌بینی کرد؛

از کجا بفهمیم واقعا خوب است؟

ماتریس درهم‌ریختگی (Confusion Matrix)

پیش‌بینی مدل روی ۱۰۰ مسافر داده‌ی تست — کلاس مثبت: «زنده ماند»

	پیش‌بینی: زنده ماند	پیش‌بینی: زنده نماند
واقعیت: زنده ماند	**۲۵** مثبت درست (TP) گفت «زنده ماند» و درست بود	**۹** منفی نادرست (FN) زنده‌مانده‌ای که از قلم افتاد
واقعیت: زنده نماند	**۸** مثبت نادرست (FP) هشدار اشتباهِ «زنده ماند»	**۵۸** منفی درست (TN) گفت «زنده نماند» و درست بود

دقت = (۲۵+۵۸) از ۱۰۰ = ٪۸۳

اما دو نوع خطای FP و FN هم‌ارزش نیستند؛ بسته به مسئله، یکی زیان‌بارتر است!

فریب دقت (Accuracy)

مدل تنبل تایتانیک: دقت ۶۲٪

برای همه بگو «زنده نماند»!
چون ۶۲٪ مسافران واقعا زنده نماندند،
بدون هیچ هوشی ۶۲٪ دقت می‌گیرد.

بیماری نادر: دقت ۹۹.۹٪

بیماری ۱ در ۱۰۰۰ نفر؛ همه را «سالم» اعلام کن!
دقت ۹۹.۹٪ می‌شود، اما حتی
یک بیمار واقعی هم پیدا نمی‌شود.

دقت به‌تنهایی کافی نیست؛ باید ببینیم مدل «کجا» اشتباه می‌کند.

پوشش (Recall): چند مورد واقعی را پیدا کردیم؟

Recall = TP ÷ (TP + FN)

از زنده‌مانده‌های واقعی، چند درصد پیدا شدند؟

۲۵ ÷ (۲۵ + ۹) ≈ ۷۴٪

پوشش بالا یعنی جامانده (FN) کم است

کجا حیاتی است؟

وقتی جاماندن زیان‌بار است:

غربال‌گری سرطان! حتی یک بیمار واقعی نباید از قلم بیفتد.

صحت (Precision): چقدر حرف مدل قابل اعتماد است؟

Precision = TP ÷ (TP + FP)

از پیش‌بینی‌های «زنده ماند»، چند درصد درست بود؟

۲۵ ÷ (۲۵ + ۸) ≈ ۷۶٪

صحت بالا یعنی هشدار اشتباه (FP) کم است

کجا حیاتی است؟

وقتی هشدار اشتباه زیان‌بار است:

فیلتر اسپم! ایمیل مهم کاری

هرگز نباید اسپم اعلام شود.

بده‌بستان صحت و پوشش

این دو معمولا با هم بالا نمی‌روند؛ سخت‌گیریِ بیشتر یعنی صحت بالاتر ولی پوشش کمتر — و برعکس.

فیلتر اسپم: صحت (Precision) مهم‌تر

فقط وقتی کاملا مطمئنی بگو «اسپم»،
حتی اگر چند اسپم از دستمان فرار کنند؛
هزینه‌ی FP بالاست.

غربال‌گری سرطان: پوشش (Recall) مهم‌تر

با کوچک‌ترین شک بگو «مشکوک»؛
آزمایش تکمیلی، هشدار اشتباه را رفع می‌کند.
هزینه‌ی FN بالاست.

معیار مناسب را «مسئله» تعیین می‌کند، نه ریاضیات!

معیار F1: آشتی‌دهنده‌ی صحت و پوشش

$$F1 = 2 \times (P \times R) \div (P + R)$$

یک عدد واحد که هر دو معیار را جدی می‌گیرد

۲ × (۰.۷۶ × ۰.۷۴) ÷ (۰.۷۶ + ۰.۷۴) ≈ ۷۵٪

میانگین هارمونیک است، نه معمولی

چرا میانگین هارمونیک؟

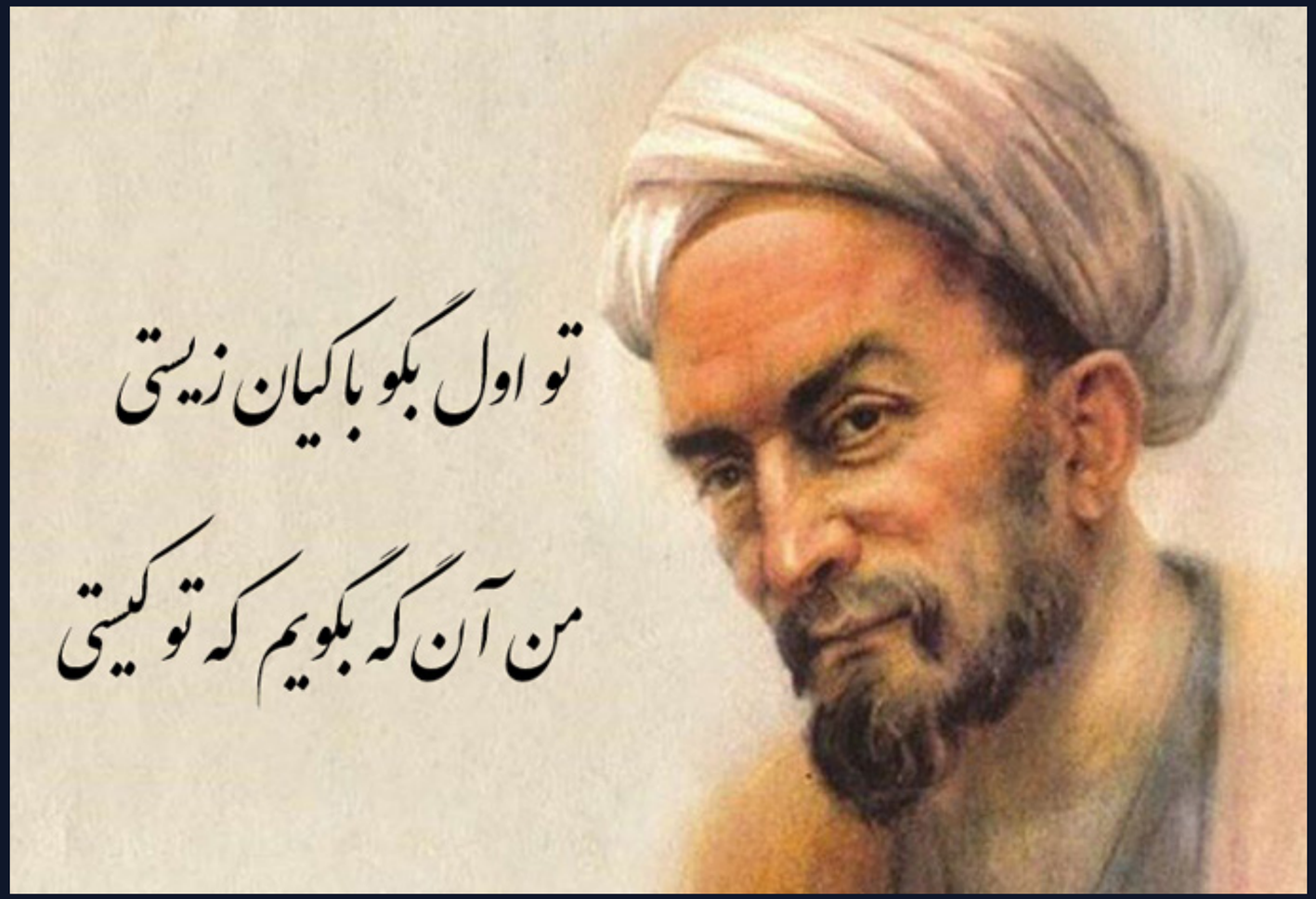
اگر یکی از دو معیار خیلی پایین باشد هم سقوط می‌کند

نمی‌شود ضعف یکی را

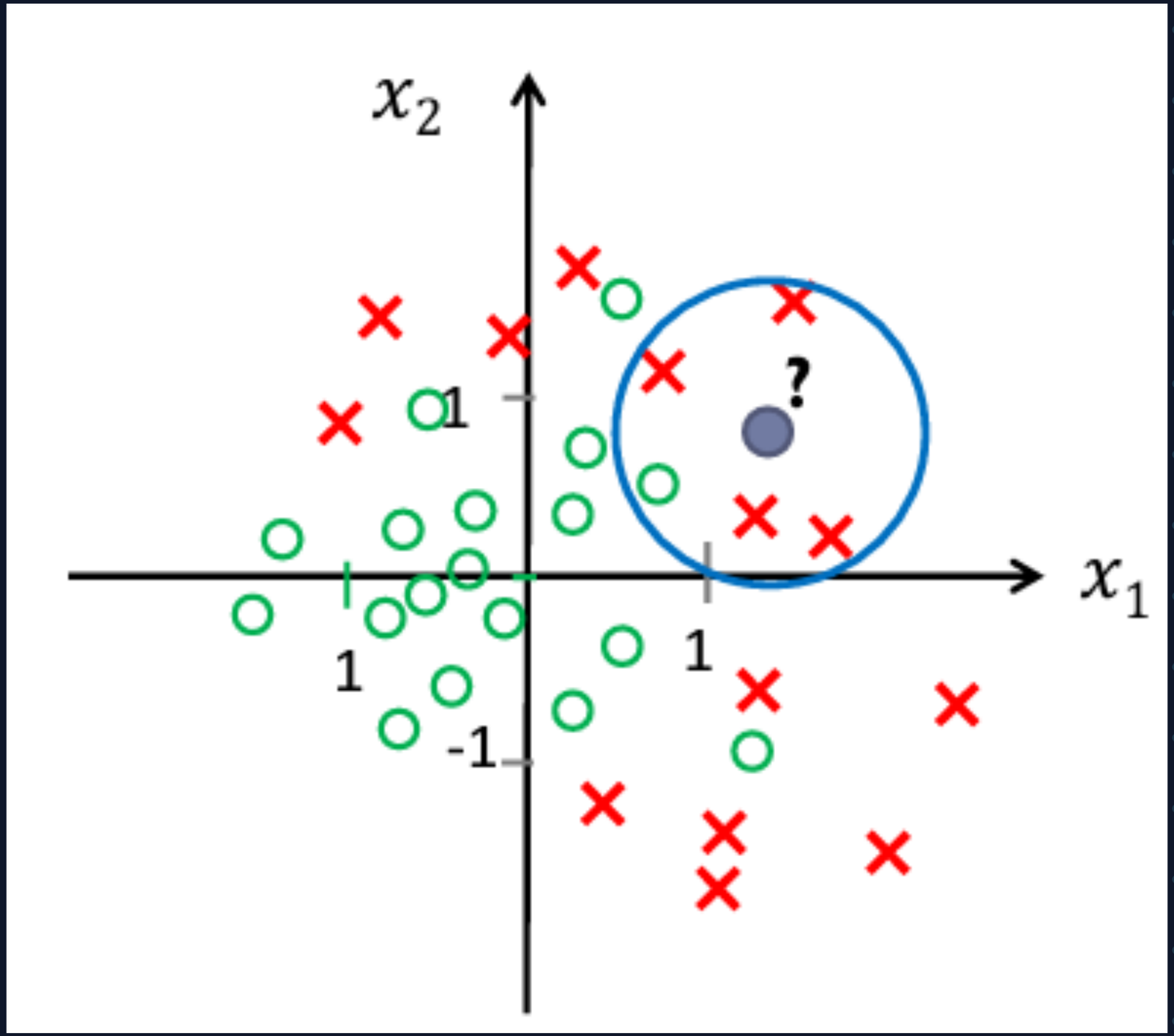
با قوت دیگری پنهان کرد.

قانون طلایی: همه‌ی معیارها را روی داده‌های ارزیابی و تست حساب می‌کنیم، نه داده‌ی آموزش!

الگوریتم K نزدیک‌ترین همسایه



الگوریتم K نزدیک‌ترین همسایه



شکل از درس یادگیری ماشین خانم دکتر مهدیه سلیمانی

نزدیک‌ترین همسایه‌ها (KNN)

- برای مسافر جدید، K مسافر شبیه به او را در داده‌ها پیدا کن (هم‌سن، هم‌کلاس، هم‌کرایه، ...)
- بین K همسایه رای‌گیری کن؛ رای اکثریت، پیش‌بینی مدل است
  - مثال با K=۵: خانم ۲۸ ساله‌ی کلاس ۲ → از ۵ مسافر شبیه او، ۴ نفر زنده مانده‌اند → زنده می‌ماند
- مرحله‌ی آموزش ندارد؛ فقط داده‌ها را نگه می‌دارد و هنگام پیش‌بینی، فاصله‌ها را می‌سنجد KNN

نقش K: چند همسایه بپرسیم؟

K = ۱

فقط نزدیک‌ترین نفر!

حساس به نویز و استثناها:

بیش‌برازش

K مناسب

رای‌گیری از جمعی معقول؛

نویز حذف می‌شود و الگوی

محلی حفظ می‌شود.

K بسیار بزرگ

رای‌گیری از کل کشتی!

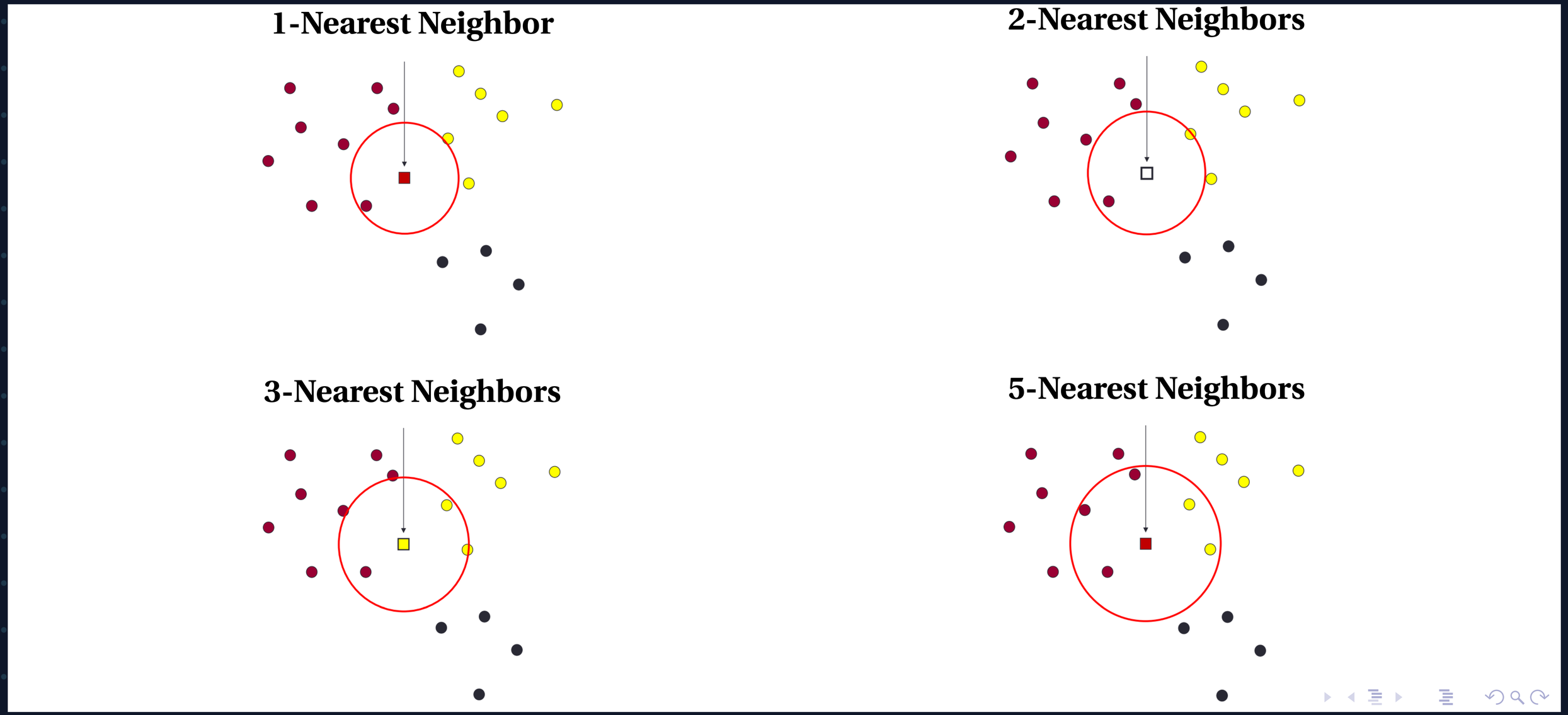
همه یک پیش‌بینی یکسان می‌گیرند:

کم‌برازش

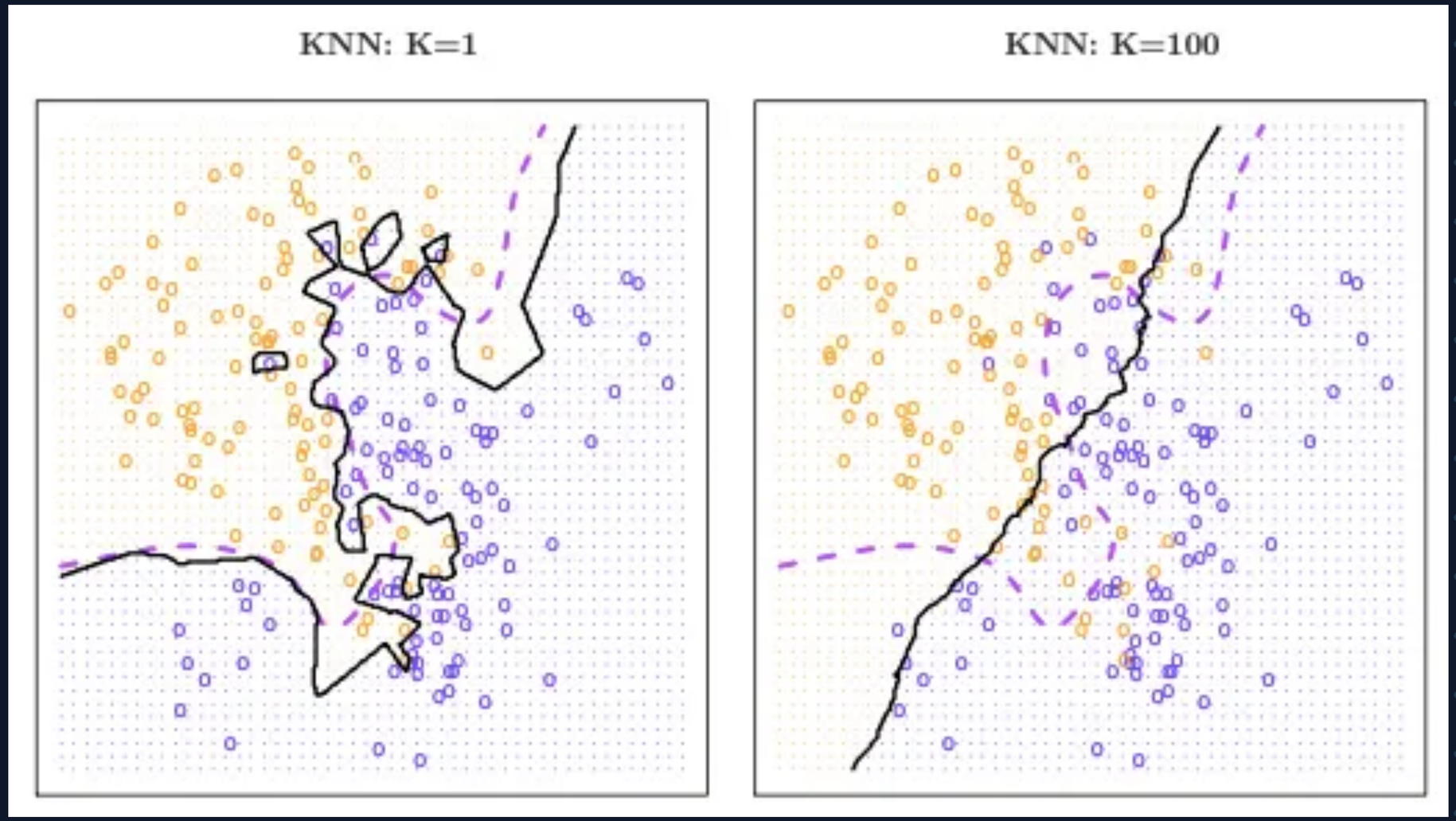
K هم یک ابرپارامتر است و می‌توانیم مقدار بهینه‌ی آن را با داده‌های Validation مشخص کنیم.

نقش K: چند همسایه بپرسیم؟

نقش K: چند همسایه بپرسیم؟

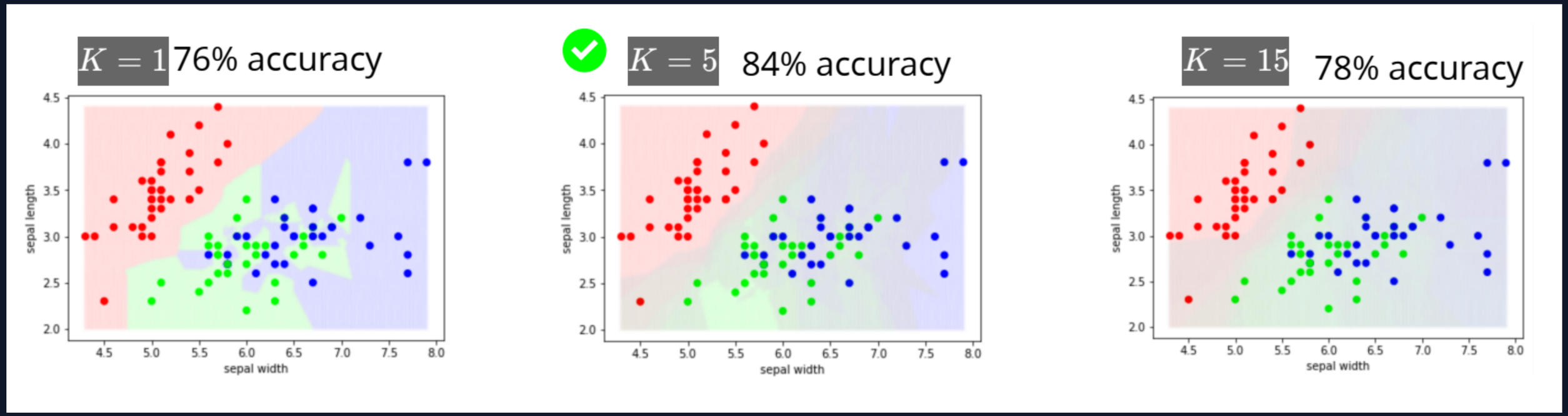


نقش K: چند همسایه بپرسیم؟

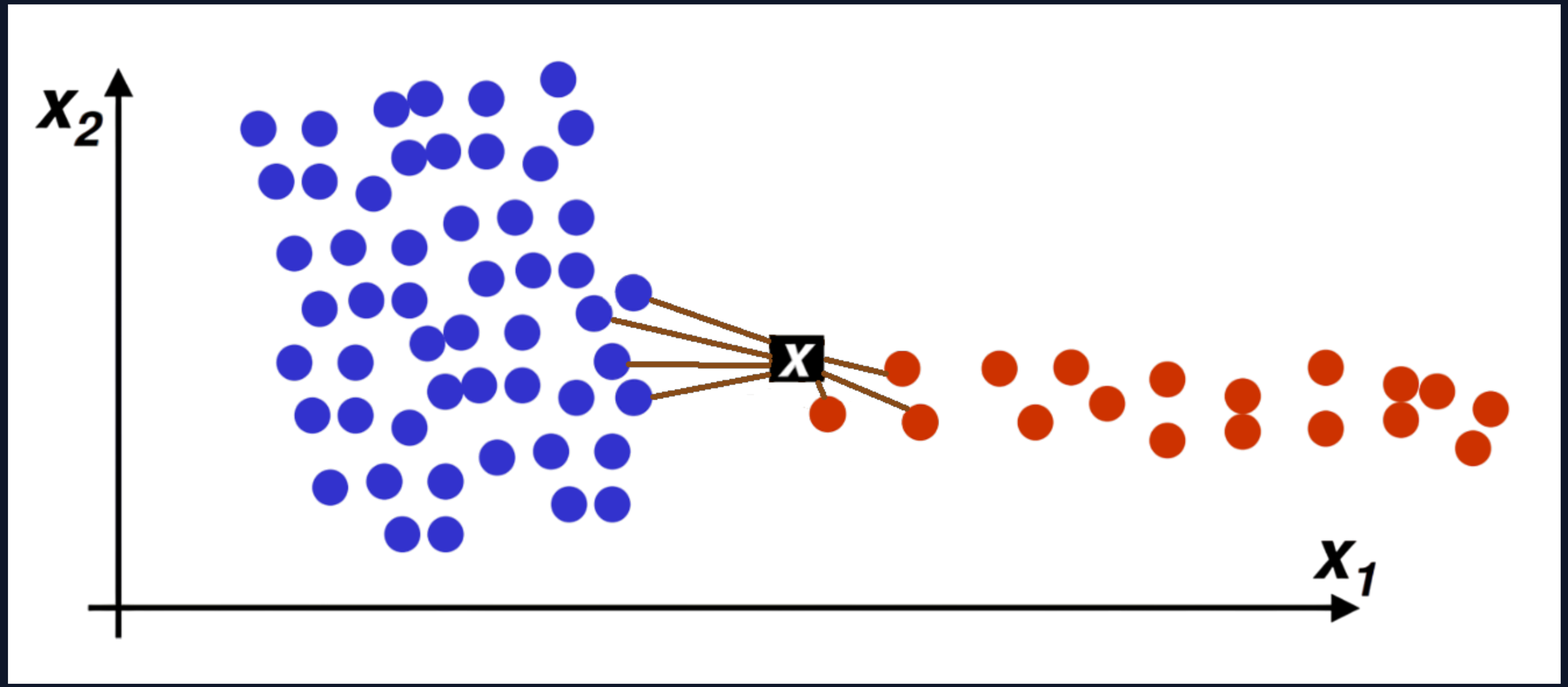


https://medium.com/@priyanshsoni761

نقش K: چند همسایه بپرسیم؟



نقش K: چند همسایه بپرسیم؟



انتخاب مقدار K

https://medium.com/@moussadoumbia_90919

دام مقیاس: فاصله را چه کسی تعیین می‌کند؟

- سن مسافران بین ۰ تا ۸۰ است؛ کرایه‌ی بلیت بین ۰ تا ۵۱۲ پوند
- در محاسبه‌ی فاصله، ۳۰ سال اختلاف سن و ۳۰ پوند اختلاف کرایه یک‌اندازه اثر دارند!
- ویژگیِ با اعداد بزرگ‌تر، فاصله را می‌بلعد و بقیه‌ی ویژگی‌ها عملا نادیده گرفته می‌شوند
- راه‌حل: هم‌مقیاس‌سازی (Scaling) — همه‌ی ویژگی‌ها به بازه‌ی مشابه، مثلا ۰ تا ۱
- درخت تصمیم به مقیاس حساس نیست، اما KNN حتما به Scaling نیاز دارد

درخت تصمیم در برابر KNN

درخت تصمیم

کاملا قابل توضیح: می‌توان درخت را «خواند»
به مقیاس ویژگی‌ها حساس نیست
خطر بیش‌برازش با عمق زیاد

نزدیک‌ترین همسایه‌ها

بدون مرحله‌ی آموزش؛ ولی پیش‌بینی کُند
حتما به هم‌مقیاس‌سازی نیاز دارد
خطر بیش‌برازش با K کوچک

دو نگاه متفاوت به یک مسئله؛ در نوت‌بوک هر دو را روی تایتانیک می‌سنجیم.

قدم بعدی: اولین مسابقه‌ی Kaggle شما

- همین مسئله‌ی تایتانیک، مسابقه‌ی شروع (Getting Started) سایت Kaggle است
- پیش‌بینی مدل خود را روی داده‌ی تست بسازید، فایل Submission را ارسال کنید و رتبه‌ی جهانی بگیرید!
  - kaggle.com/c/titanic
- چالش هفته: با تغییر ویژگی‌ها و ابرپارامترها، دقت بالای ۸۰٪ بگیرید