یادگیری با نظارت

یادگیری مبتنی بر نمونه (Instance-Based Learning)

المپیاد هوش مصنوعی - یادگیری ماشین

مسئله‌ای که در این بخش می‌خواهیم به آن بپردازیم این است: n داده‌ی آموزشی و m کلاس داریم و هر یک از داده‌ها متعلّق به یکی از کلاس‌هاست. حال می‌خواهیم برای یک داده‌ی جدید مشخّص کنیم با چه احتمالی متعلّق به هر یک از کلاس‌هاست.

مثال

اگر ۳ کلاس داشته باشیم ممکن است با دیدن یک داده‌ی جدید بگوییم که این داده با احتمال ٪۸۰ برای کلاس ۱، با احتمال ٪۱۵ برای کلاس ۲، و با احتمال ٪۵ برای کلاس ۳ است.

به عبارت دیگر، می‌خواهیم یک (یا چند) تابع چگالی احتمال را از روی داده‌های آموزشی تخمین بزنیم.

- **تخمین توابع چگالی دلخواه**
  - توابع چگالی پارامتری (مثل گوسی) معمولاً نمی‌توانند بر چگالی‌های پیچیده‌ی دنیای واقعی منطبق شوند (مثلاً داده‌هایی که چند قله یا مُد دارند).
  - روش‌های غیرپارامتری هیچ پیش‌فرضی درباره شکل نمودار (مثلاً زنگوله‌ای بودن) ندارند.
- روش‌های غیرپارامتری برای طبقه‌بندی:
  - روش‌های Generative: تخمین $p(x|C_i)$ (چگالی داده‌ها در هر کلاس).
  - روش‌های Discriminative: تخمین مستقیم $p(C_i|x)$ (احتمال کلاس بودن به شرط داده).

روش‌های پارامتری:

- فرض: داده‌ها از یک فرمول خاص (مثلاً خطی یا گوسی) پیروی می‌کنند.
- کار ما: پیدا کردن پارامترهای آن فرمول (مثلاً شیب خط یا میانگین).
- پس از یادگیری، داده‌های آموزشی را دور می‌ریزیم.

روش‌های غیرپارامتری:

- فرض: هیچ فرضی نداریم! داده‌ها هر شکلی می‌توانند باشند.
- کار ما: نگه داشتن تمام داده‌ها.
- فاز "آموزش" خاصی نداریم؛ یادگیری در لحظه‌ی پرسش (تست) انجام می‌شود.

در روش‌های بدون پارامتر، منطق ساده است:

«بگو دوستانت کیستند تا بگویم کیستی!»

برای تصمیم‌گیری درباره یک داده‌ی جدید (x)، به داده‌های آموزشیِ نزدیک به آن نگاه می‌کنیم. دو راه برای تعریف «نزدیکی» داریم:

1. **روش پنجره پارزن (Parzen Window):** شعاع ثابتی (h) را دور x خط می‌کشیم. هر کسی داخل این شعاع بود نظر می‌دهد. (تعداد همسایه‌ها متغیر است).
2. **روش k نزدیک‌ترین همسایه (KNN):** دقیقاً k نفر از نزدیک‌ترین افراد را پیدا می‌کنیم، فارغ از اینکه چقدر دور باشند. (شعاع همسایگی متغیر است).

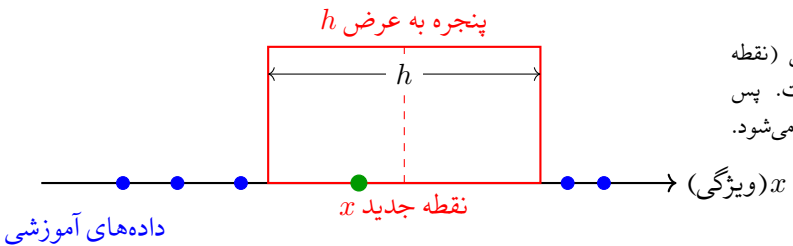
قبل از تعریف ریاضی، بیایید تصور کنیم یک پنجره (یا جعبه) داریم:

- این جعبه را دقیقاً روی داده‌ی جدید x قرار می‌دهیم.
- عرض این جعبه برابر با h است.
- هر داده‌ی آموزشی که داخل این جعبه بیفتد، یک «رای» یا «اثر» دارد.
- مجموع این رای‌ها، چگالی (تراکم) را در نقطه x می‌سازد.

سؤال: شکل این جعبه چگونه باشد؟

- ساده‌ترین حالت: یک مستطیل تخت (ابرمکعب). هر کس داخل بود اثر ۱ دارد، هرکس بیرون بود اثر ۰.
- حالت پیشرفته (کرنل گوسی): تاثیر افراد نزدیک‌تر بیشتر از افراد دورتر است (شبیه تپه).

- تابع پنجره (کرنل) ساده: اگر فاصله داده تا مرکز کمتر از $h/2$ بود، مقدار ۱ وگرنه ۰.
- در ریاضیات، این را با یک تابع نشان می‌دهیم که اگر ورودی‌اش کوچک باشد ۱ می‌دهد.



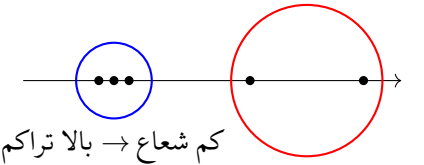
در این مثال، تنها یک داده آموزشی (نقطه سبز) درون پنجره‌ی دور افتاده است. پس چگالی در اینجا کم تخمین زده می‌شود.

شکل ۱: نمایش هندسی پنجره پارزن یک بعدی (ابرمکعب)

پارامتر h (عرض پنجره) نقش کلیدی دارد. تصور کنید می‌خواهید پستی و بلندی‌های یک تپه (تابع چگالی) را پیدا کنید:

- **اگر h خیلی بزرگ باشد (تار دیدن):**
  - پنجره آنقدر بزرگ است که همه‌ی داده‌ها را با هم قاطی می‌کند. جزئیات و قله‌های کوچک از بین می‌روند.
  - نتیجه: نمودار خیلی صاف (Smooth) می‌شود.
  - به زبان علمی: **بایاس زیاد** (چون واقعیت را ساده کرده‌ایم) و **واریانس کم** (با تغییر داده‌ها شکل کلی عوض نمی‌شود).
- **اگر h خیلی کوچک باشد (جزئی‌نگری افراطی):**
  - پنجره فقط روی تک‌تک نقاط متمرکز می‌شود. نمودار پر از تیغه‌های باریک می‌شود.
  - نتیجه: نویزهای داده به عنوان الگو شناخته می‌شوند.
  - به زبان علمی: **بایاس کم** (دقیق روی داده‌ها) اما **واریانس زیاد** (حساس به نویز).

- در روش قبل h ثابت بود، اما اینجا k (تعداد همسایه) ثابت است.
- ما پنجره را آنقدر بزرگ می‌کنیم تا دقیقا k همسایه را در بر بگیرد.
- جایی که داده‌ها متراکم هستند، پنجره کوچک می‌شود (دقت بالا).
- جایی که داده‌ها خلوت هستند، پنجره بزرگ می‌شود (دقت پایین).

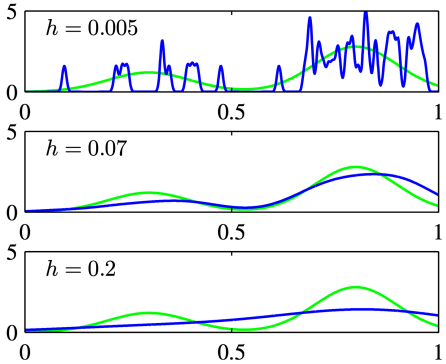


انتخاب تعداد همسایگان (k) مثل انتخاب تعداد داوران در یک مسابقه است:

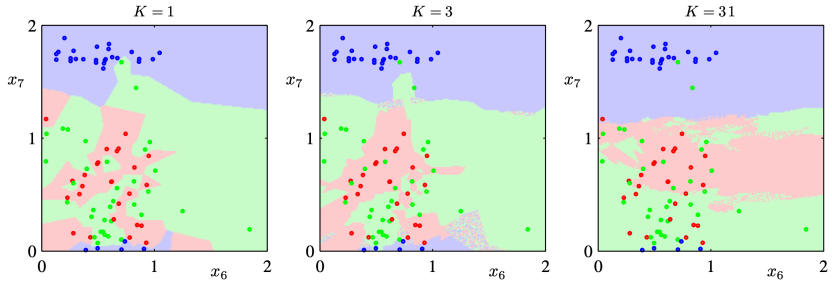
- **اگر $k = 1$ باشد (واریانس زیاد):**
  - شما فقط به حرف نزدیک‌ترین همسایه گوش می‌دهید.
  - اگر آن یک نفر اشتباه کرده باشد (نویز باشد)، شما هم اشتباه می‌کنید. تصمیم‌گیری بسیار متزلزل و ناپایدار است.
- **اگر k خیلی بزرگ باشد (بایاس زیاد):**
  - شما از کل جمعیت شهر (حتی کسانی که خیلی دورند و ربطی به شما ندارند) نظر می‌خواهید.
  - تفاوت‌های محلی از بین می‌رود و نظر اکثریت کل جامعه غلبه می‌کند. مرزهای بین کلاس‌ها محو می‌شود.
- مقدار تجربی خوب: معمولاً $k = \sqrt{n}$ پیشنهاد می‌شود.

- **قدرت:** با داده‌ی کافی، هر تابع پیچیده‌ای را یاد می‌گیرند.
- **هزینه:** "تنبلی" در یادگیری باعث می‌شود در زمان تست زحمت زیادی بکشیم (محاسبات سنگین برای پیدا کردن فاصله با همه).
- **نفرین ابعاد:** در ابعاد بالا (تعداد ویژگی‌های زیاد)، مفهوم «فاصله» و «همسایگی» خراب می‌شود و به تعداد نمایی داده نیاز داریم.

در شکل زیر تاثیر اندازه پنجره را می‌بینیم. به تفاوت بین "صاف بودن بیش از حد" و "تیز بودن بیش از حد" دقت کنید.



شکل ۲: تاثیر h در روش پنجره‌ی پارزن



شکل ۳: مقادیر کوچک k باعث مرزهای دندانه‌دار (واریانس بالا) می‌شوند.

فاصله: چگونه شباهت را بسنجیم؟

تمام این روش‌ها بر اساس "نزدیکی" کار می‌کنند. اما نزدیکی یعنی چه؟

- معمول‌ترین روش: **فاصله‌ی اقلیدسی** (خط کشی).

$$d(x, x') = \sqrt{\sum_{i=1}^{d} (x_i - x'_i)^2}$$

- مشکل: اگر یک ویژگی عدد بزرگی باشد (مثلا حقوق به ریال) و دیگری کوچک (مثلا سن)، ویژگی بزرگ‌تر بر فاصله مسلط می‌شود.
- راه حل: استفاده از **فاصله‌ی وزن‌دار** یا نرمال‌سازی داده‌ها قبل از شروع کار.

$$d_w(x, x') = \sqrt{\sum_{i=1}^{d} w_i (x_i - x'_i)^2}$$

شما میانگین ۵ نفری هستید که بیشترین وقت را با آنها می‌گذرانید.
(جیم ران - تفسیر اجتماعی از الگوریتم KNN)

- یادگیری مبتنی بر نمونه = حافظه‌ی قوی + سنجش شباهت.
- بدون فرض اولیه (Non-parametric) = انعطاف‌پذیری بالا.
- هزینه = کندی در زمان اجرا (چون مدل فشرده‌ای نداریم).