یادگیری ماشین

دکتر محمد حسین رهبان

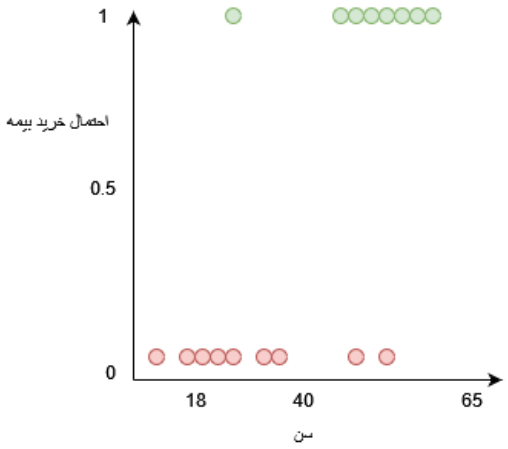
AI CONTEST rayan

مباحث این جلسه

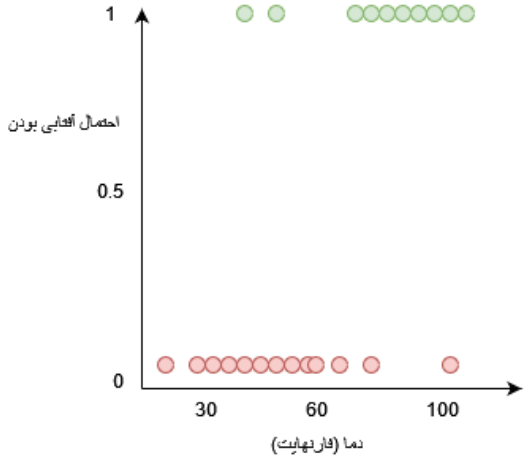
- لاجیستیک رگرشن (Logistic Regression)
- انتخاب مدل (Model Selection)
- تقسیم داده (Data Split)
- سوگیری قیاسی (Inductive Bias)
- تعمیم‌پذیری (Generalization)

مسئله

چگونه مسائل زیر را مدل سازی کنیم؟



پیش‌بینی خرید بیمه

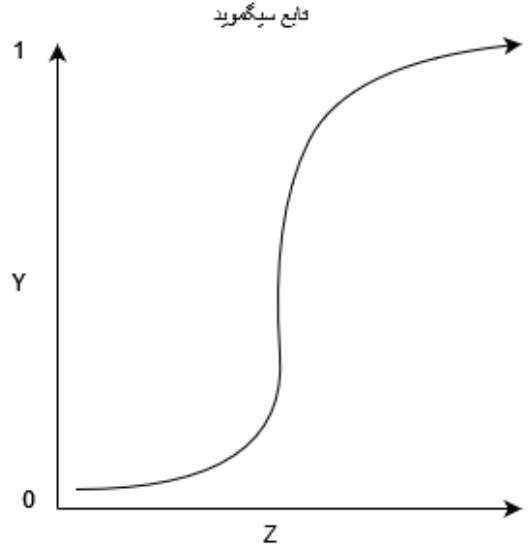


پیش‌بینی آب و هوا

Logistic Regression

- **تعریف** : استفاده از تابع لاجستیک برای مدل‌سازی یک متغیر وابسته باینری (خرید بیمه) از روی یک یا چند متغیر مستقل (سن)
- **کاربرد** : مسائل دسته‌بندی دو کلاسه (مانند تشخیص هرزنامه، تشخیص بیماری و غیره.)
- **خروجی** : احتمال تعلق به یک کلاس خاص

Sigmoid Function



تابع Sigmoid

$$y = \sigma(z) = \frac{1}{1 + e^{-z}}$$

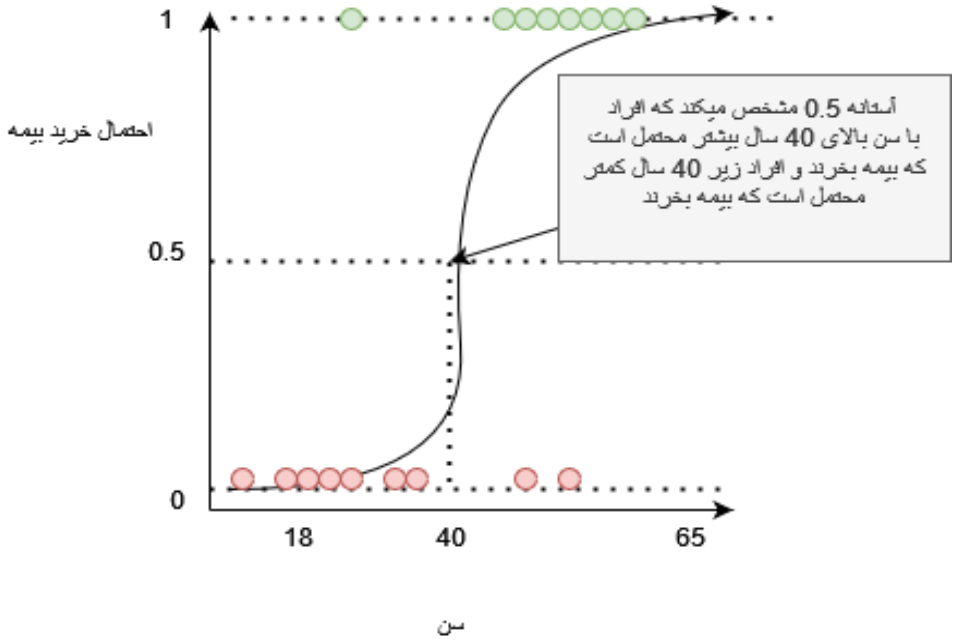
یک متغیر مستقل

$$z = \beta_0 + \beta_1 x_1$$

چند متغیر مستقل

$$z = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

مسئله خرید بیمه



مدل‌سازی مسئله پیش‌بینی خرید بیمه با لاجستیک رگرشن

مثال تعاملی: لجستیک رگرشن

مسئله

در مسأله پیش‌بینی قیمت خانه بر اساس ویژگی‌هایی نظیر محله، متراژ، تعداد اتاق خواب و ... با چه چالش‌هایی در انتخاب مدل روبرو هستیم؟

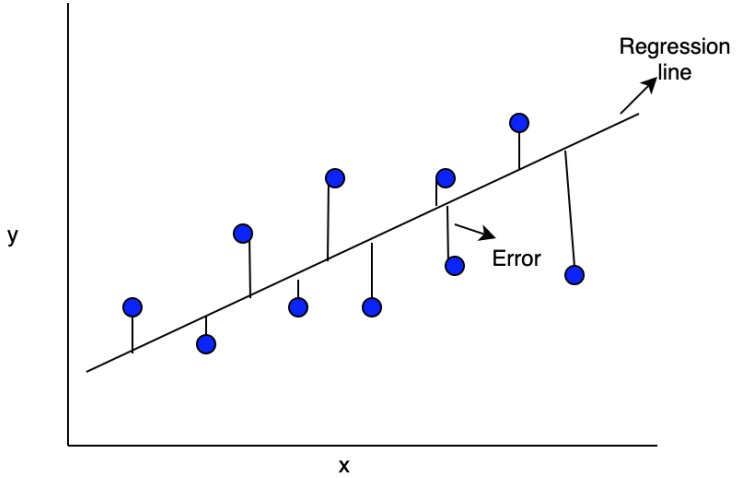
دقت مدل در پیش‌بینی

عملکرد روی داده‌های جدید

اعتمادپذیری و تفسیرپذیری

Inductive Bias

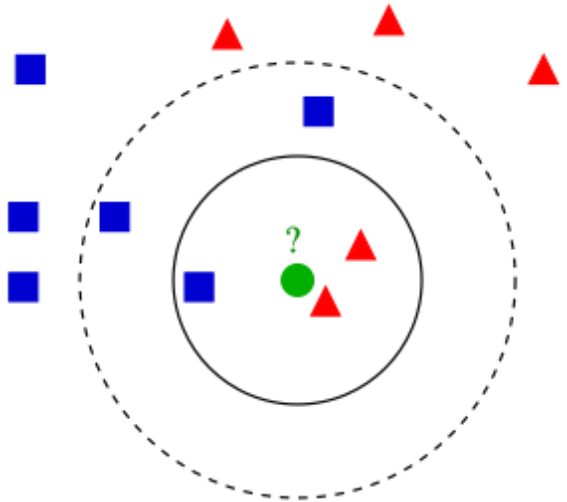
تعریف : فرضیاتی که مدل برای تعمیم از داده آموزش به داده تست استفاده می‌کند.



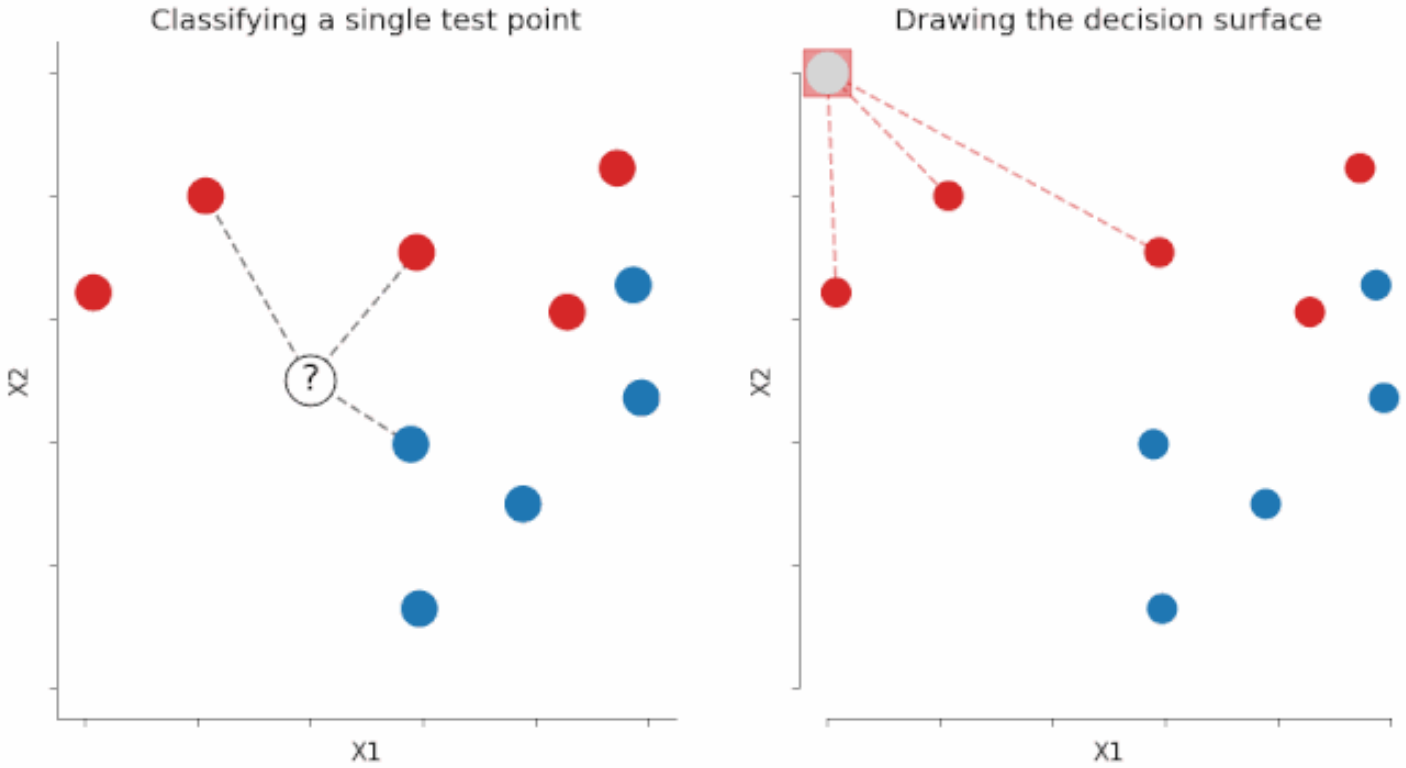
مثال : مدل رگرشن خطی: وجود رابطه خطی بین ورودی و خروجی

یک روش ساده دیگر

دسته‌بند k-NN : برای داده x یافتن k نزدیک‌ترین داده آموزشی به آن
کلاسی اکثریت در بین همسایه‌ها به x نسبت داده می‌شود



*k*NN classifier | $k = 3$



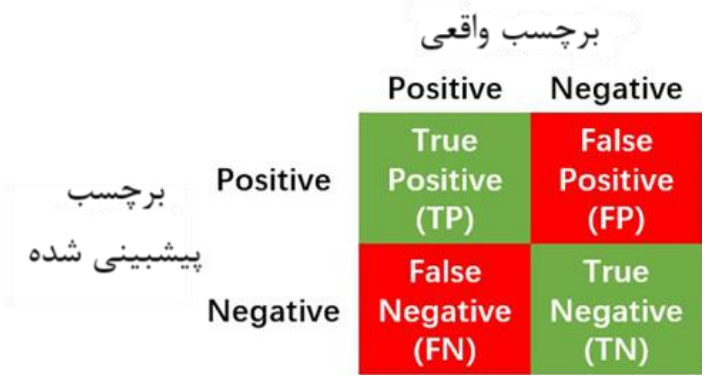
انتخاب مدل (Model Selection)

k-NN بهتره یا Logistic Regression؟

برای این منظور چه فاکتورهایی را باید در نظر بگیریم؟

- **سوگیری قیاسی** : فرضیات پشت کدام مدل با مسئله سازگار است؟
- **پیچیدگی مدل** : تقابل بین سادگی مدل و عملکرد مدل
- **عملکرد مدل** : Accuracy, Precision, Recall
- **تفسیرپذیری** : فهمیدن اینکه مدل بر چه اساسی پیش‌بینی می‌کند

متریک‌های ارزیابی



$$Precision = \frac{\sum TP}{\sum TP + FP}$$

$$Recall = \frac{\sum TP}{\sum TP + FN}$$

$$Accuracy = \frac{\sum TP + TN}{\sum TP + FP + TN + FN}$$

$$Sepecificity = \frac{\sum TN}{\sum TN + FP}$$

مدلی داریم که برای تشخیص سرطان به دقت 99 درصد رسیده است.
نظر شما در مورد این مدل چیست؟

تقسیم‌بندی داده‌ها

در یادگیری ماشین داده‌ها به سه قسمت زیر تقسیم‌بندی می‌شوند:

آموزش

Train

ارزیابی

Validation

تست

Test

آموزش، ارزیابی و تست

- **داده آموزش** : برای آموزش مدل استفاده می‌شود
- **داده‌ی ارزیابی** : برای تیون کردن هایپرپارامتر و انتخاب مدل استفاده می‌شود
- **داده‌ی تست** : داده‌ای یکبار مصرف که برای ارزیابی نهایی مدل استفاده می‌شود و عملکرد مدل نهایتا توسط این داده گزارش می‌شود
- **نحوه جداسازی متداول** : 0.8 Train ,0.1 noitadilaV ,0.1 tseT

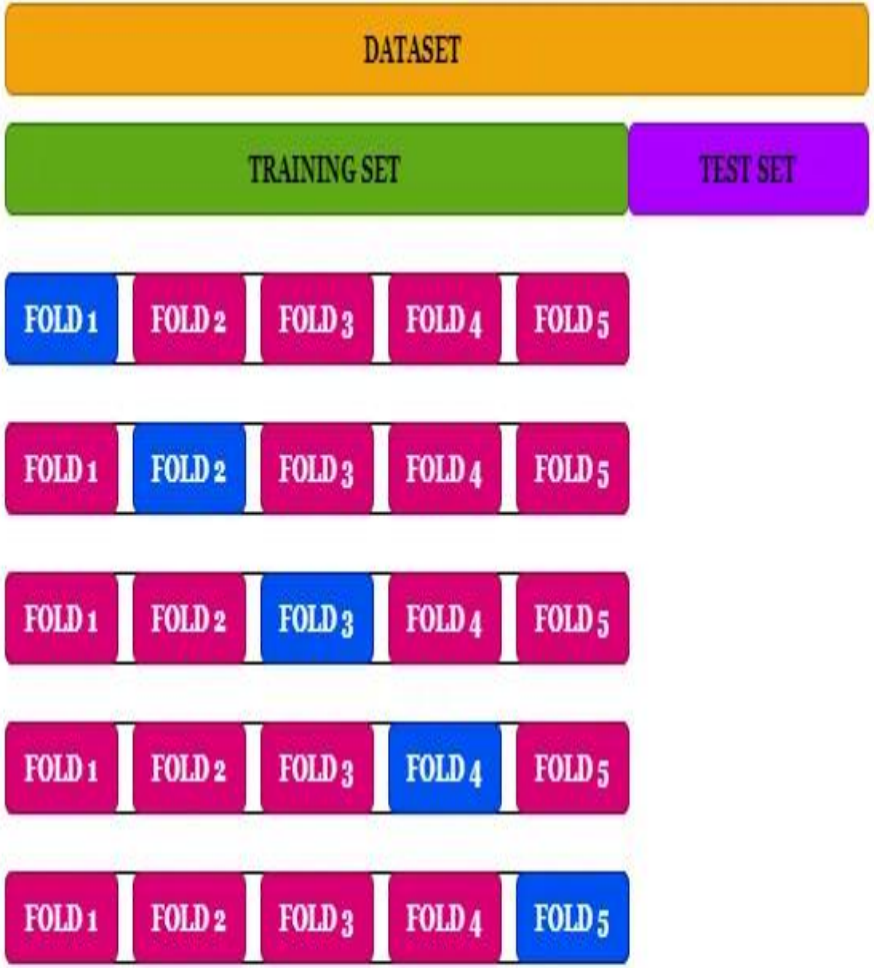
اهمیت جداسازی داده

- اطمینان از اینکه مدل داده‌ها را حفظ نکرده (overfitting)
- انتخاب بهترین مدل و تیون کردن هایپرپارامترها با استفاده از داده ارزیابی بدون نشت اطلاعات از داده تست
- عملکرد منصفانه و بدون غرض روی داده‌های واقعی دیده نشده با استفاده از داده تست.

Cross Validation

تعریف : تکنیکی برای ارزیابی عملکرد مدل با استفاده از چند دسته‌ای کردن داده‌های آموزش و ارزیابی

هدف : بدست آوردن تقریب دقیق‌تر از عملکرد مدل



K-Fold Cross Validation

تعریف : تقسیم داده به k زیر دسته، سپس k بار آموزش مدل و ارزیابی روی زیر دسته‌های متفاوت

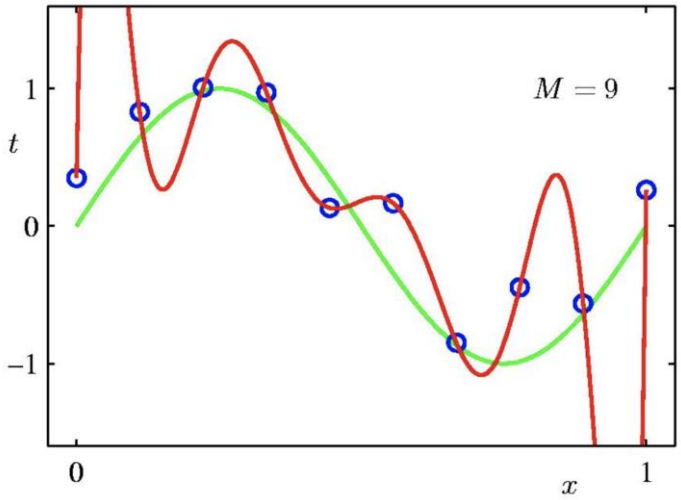
تعمیم‌پذیری

تعمیم‌پذیری (Generalization) : توانایی یک مدل برای عملکرد خوب بر روی داده‌های دیده نشده

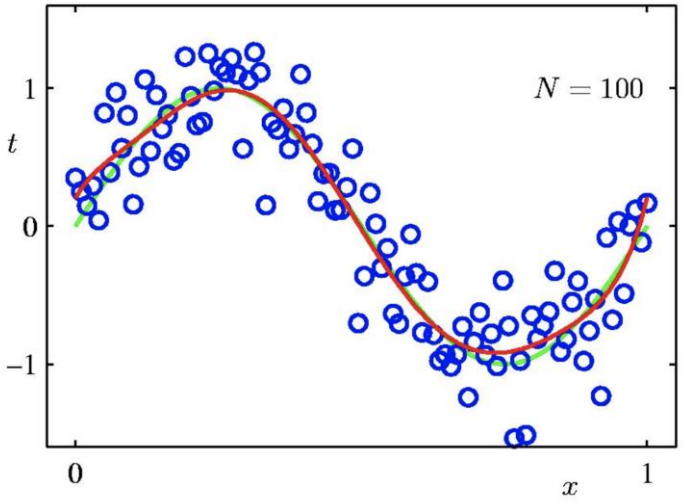
فاکتورهای مهم :

- **پیچیدگی مدل** : تقابل بین پیچیدگی و عملکرد که مدل‌های ساده underfit می‌شوند و مدل‌های پیچیده overfit
- **اندازه داده آموزش** : داده بیشتر عموما به تعمیم‌پذیری بیشتر منجر می‌شود
- **مهندسی داده** : کیفیت و مرتبط بودن ویژگی‌های موجود در داده

تعمیم‌پذیری



ساده‌تر کردن مدل



افزایش حجم داده‌ها

ارزیابی تعمیم‌پذیری

چگونه تعمیم‌پذیری را ارزیابی کنیم؟

ارزیابی روی داده ارزیابی

ارزیابی روی داده تست

استفاده از Cross Validation
برای تخمین تعمیم‌پذیری

Machine Learning Pipeline