

# Classification

## داده های گسسته

در آخرین ماژول، ما یک مدل رگرسیون خطی برای پیش‌بینی یک مقدار پیوسته، مقدار متوسط خانه در بوستون ساختیم. در این ماژول، ما مشکلات طبقه‌بندی را بررسی می‌کنیم که وظیفه آنها پیش‌بینی یک مقدار گسسته است.

داده‌های گسسته فقط می‌توانند مقادیر خاصی داشته باشند، در حالی که داده‌های پیوسته می‌توانند هر مقداری را بگیرند.

- نمونه هایی از مشکلات طبقه بندی که شامل مقادیر داده های گسسته است عبارتند از:
- برای پیش بینی خوش خیم یا بدخیم بودن سرطان سینه با توجه به مجموعه ای از ویژگی ها
  - طبقه بندی یک تصویر به عنوان حاوی گربه یا سگ یا اسب
  - برای پیش بینی اینکه آیا یک ایمیل از یک آدرس ایمیل داده شده هرزنامه است یا خیر

در هر یک از مثال‌ها، برچسب‌ها به صورت دسته‌بندی هستند و تعداد محدودی از کلاس‌ها را نشان می‌دهند.

## Binary and Multi-class Classification

دو نوع طبقه بندی وجود دارد: باینری و چند کلاسه. اگر دو دسته برای پیش بینی وجود داشته باشد، این یک مشکل طبقه بندی باینری است، به عنوان مثال، یک تومور خوش خیم یا بدخیم. وقتی بیش از دو کلاس وجود دارد، وظیفه یک مشکل چند طبقه بندی است. به عنوان مثال، طبقه بندی گونه های زنبق، که می تواند *versicolor*، *virginica* یا *setosa* باشد، بر اساس ویژگی های کاسبرگ و گلبرگ آنها.

الگوریتم های رایج برای طبقه بندی عبارتند از رگرسیون لجستیک،  $k$  نزدیکترین همسایه، درخت تصمیم، خلیج ساده، ماشین های بردار پشتیبان، شبکه های عصبی و غیره. در اینجا نحوه استفاده از  $k$  نزدیکترین همسایه برای طبقه بندی گونه های عنبیه را یاد خواهیم گرفت.