یا حق

نام: حسین علمردانی صومعه

شماره دانشجویی: ۹۶۳۹۲۶۶۱۱

استاد: استاد ماهی

دانشگاه: پیام‌نور مرکز-تبریز

موضوع: الگوریتم K-NN

درس: اصول مدیریت و برنامه ریزی راهبردی فناوری اطلاعات

مقدمه

K-نزدیک‌ترین همسایگی (k-Nearest Neighbors) یک روش ناپارامتری است که در داده‌کاوی، یادگیری ماشین و تشخیص الگو مورد استفاده قرار می‌گیرد. بر اساس آمارهای ارائه شده در وب‌سایت kdnuggets الگوریتم K-نزدیک‌ترین همسایگی یکی از ده الگوریتمی است که بیشترین استفاده را در پروژه‌های گوناگون یادگیری ماشین و داده‌کاوی، هم در صنعت و هم در دانشگاه داشته است.

فهرست مطالب

یکی از دلایل اصلی پرکاربرد بودن الگوریتم‌های طبقه‌بندی (Classification) آن است که «تصمیم‌گیری» یکی از چالش‌های اساسی موجود در اغلب پروژه‌های تحلیلی است. برای مثال، تصمیم‌گیری درباره اینکه آیا مشتری X پتانسیل لازم برای مورد هدف قرار داده شدن در کارزارهای دیجیتال یک کسب‌وکار را دارد یا خیر و یا اینکه آیا یک مشتری وفادار است یا نه از جمله مسائل تصمیم‌گیری به حساب می‌آیند که در فرآیند تحلیل قصد پاسخ‌دهی به آن‌ها وجود دارد. نتایج این تحلیل‌ها بسیار تأمل‌برانگیز هستند و به‌طور مستقیم به پیاده‌سازی نقشه راه در یک سازمان یا کسب‌وکار کمک می‌کنند. در این نوشتار، به یکی از روش‌های پرکاربرد

طبقه‌بندی، یعنی روش k-نزدیک‌ترین همسایگی پرداخته شده و تمرکز آن بر چگونگی کار کردن الگوریتم و تأثیر پارامترهای ورودی بر خروجی و پیش‌بینی است.

چه زمانی باید از الگوریتم k-نزدیک‌ترین همسایگی استفاده کرد؟

الگوریتم k-نزدیک‌ترین همسایگی برای مسائل طبقه‌بندی و رگرسیون قابل استفاده است. اگرچه، در اغلب مواقع از آن برای مسائل طبقه‌بندی استفاده می‌شود. برای ارزیابی هر روشی به طور کلی به سه جنبه مهم آن توجه می‌شود:

۱. سهولت تفسیر خروجی‌ها
۲. زمان محاسبه
۳. قدرت پیش‌بینی

در جدول ۱ الگوریتم نزدیک‌ترین همسایگی با الگوریتم‌های «رگرسیون لجستیک»، «CART »و «جنگل‌های تصادفی» (random forests) مقایسه شده است. همان‌گونه که از جدول مشخص است، الگوریتم k-نزدیک‌ترین همسایگی بر اساس جنبه‌های بیان شده در بالا، نسبت به دیگر الگوریتم‌های موجود در جایگاه مناسبی قرار دارد. این الگوریتم اغلب به دلیل سهولت تفسیر نتایج و زمان محاسبه پایین مورد استفاده قرار می‌گیرد.

	Logistic Regression	CART	Random Forest	KNN