یادگیری ماشین

Machine Learning

۱) مقاله با عنوان "**k-Nearest Neighbour Classifiers - A Tutorial**" را مطالعه کنید و به سوالات زیر پاسخ دهید

الف) استفاده از مدل Vote در چه مواقعی مناسب است؟

ب) از همبستگی Spearman در چه شرایطی استفاده می‌شود؟

پ) نقاط ضعف معیار Kullback-Leibler چیست؟

ت) از مهمترین نقاط ضعف kNN حساسیت آن به چیست؟

ث) برای بهبود سرعت اجرای kNN می‌توان از راهکارهای Kd Tree یا Ball Tree استفاده کرد. این دو روش را با یکدیگر مقایسه کنید و به دلخواه نحوه اجرای یکی از این دو روش را توضیح دهید.

۲) برای داده مشخص شده در این تمرین (طبقه‌بندی متن فارسی و عربی: Ar-Fa)، مطلوبست

الف) بردار ویژگی را به صورت **BoW دودویی** ایجاد کنید، برای این کار می‌توان از کاراکترهای موجود در متن هر نمونه استفاده کرد، به این صورت که برای هر متن نمونه، یک بردار ویژگی داریم که مقادیر این بردار برابر با وجود یا عدم وجود کاراکتر مورد نظر در آن نمونه متن است. مثلا اگر فرض کنیم که بردار ویژگی به صورت (A, B, C, D, E) تعریف شود، آنگاه برای عبارتی به صورت ABBCAAB بردار ویژگی به صورت (1, 1, 1, 0, 0) خواهد بود.

ب) بردار ویژگی را به صورت **BoW وزن‌دار** ایجاد کنید، برای این کار مشابه BoW باینری عمل می‌شود، با این تفاوت که مقدار درایه بردار مورد نظر برابر فراوانی کاراکترهای موجود در آن متن است. یعنی بردار ویژگی برای مثال مذکور به صورت (3, 3, 1, 0, 0) خواهد بود.

پ) بردار ویژگی را به صورت **BoW نرمال شده طول** ایجاد کنید، برای این کار مقادیر بدست آمده برای BoW وزن‌دار هر بردار را بر مجموعه وزنی بردار (اندازه متن) تقسیم می‌کنیم، یعنی برای بردار ویژگی داریم (3/7, 3/7, 1/7, 0, 0).

ت) بردار ویژگی را به صورت **BoW نرمال شده نمره-زد** ایجاد کنید، برای این کار مقادیر بدست آمده برای هر درایه BoW وزن‌دار را متناسب با میانگین و انحراف معیار مجموع مقادیر متناظر با آن درایه نرمال کنید. یعنی برای بردار ویژگی ممکن است داشته باشیم (1.45, 2.32, 0.74, -1.1, 0.02)

ث) حال با استفاده از هر یک از بردار ویژگی‌های بدست آمده از بخش‌های قبلی، عملکرد الگوریتم kNN را براساس خطا (تعداد پیشبینی‌های نادرست به روی کل تعداد پیش‌بینی‌ها برای نمونه‌های ۱۰، ۲۰، ۳۰، ۴۰، ۵۰، ۶۰، ۷۰، ۸۰، ۹۰ و ۱۰۰) به ازای k=1,3,5 و حداقل ۲ معیار فاصله مختلف (مثلا Cosine, Euclidian) در جدول زیر گزارش کنید.

ج) (**اختیاری و امتیازی**): از چه روش‌های دیگری می‌توان بردار ویژگی را ایجاد کرد؟ و همچنین، از چه معیارهای دیگری می‌توان برای محاسبه فاصله استفاده کرد؟ علاوه بر توضیح هر یک، در جدول نتایج شبیه‌سازی را گزارش کنید.

	K=1		K=2		K=3	
معیار فاصله / ویژگی	Euclidian	Cosine	Euclidian	Cosine	Euclidian	Cosine
BoW دودویی						
BoW وزندار						
BoW نرمال شده طول						
BoW نرمال شده نمره-زد						

توجه: هر گونه فعالیت اضافی و موثر از جانب خودتان را در گزارش مشخص کنید تا مورد ارزیابی قرار گیرد.