به نام خدا

دانشکده علوم و فنون نوین

مباحث ویژه
یادگیری ماشین و شبکه‌های عصبی مصنوعی

نیم‌سال دوم ۱۳۹۸-۱۳۹۷
تمرین شماره ۱

تاریخ تحویل: ۱۳۹۷/۱۲/۲۷

---

۱. (۱۰٪) [**مطالعه و موضوع پروژه**] شبکه‌های عصبی مصنوعی در کاربردهای مختلفی و به شیوه‌های متنوعی به کار گرفته شده‌اند. هدف این تمرین، مروری بر کاربردها و نحوه به‌کارگیری آنها در زمینه مربوط به رشته شماست. برای این کار، هر دانشجو، حداقل دو کاربرد مختلف را که در مقالات علمی آمده‌اند، مورد مطالعه قرار داده و گزارش آن را به صورت مکتوب (الکترونیکی) ارسال کند. پاسخ خود را به صورت یک گزارش علمی تهیه کرده و در آن، منابع علمی مورد استفاده (با تاکید بر ژورنال‌ها و پایان‌نامه‌ها) را به صورت استاندارد بیان کنید. منابع مورد استفاده خود را نیز به همراه گزارش تحویل دهید. از میان موضوع‌های مطالعه شده، کدامیک را به عنوان موضوع پروژه برمی‌گزینید؟

---

۲. (۱۰٪) [**مطالعه و تحلیل**] تفاوت‌های میان یادگیری ماشین، داده‌کاوی و بازشناسی الگو را ذکر کنید.

---

۳. (۷۰٪) [**پیاده‌سازی: یک دسته‌بند ساده و معیارهای ارزیابی**] در این مسئله شما یک دسته‌بند ساده را به منظور تشخیص سه زبان فارسی، عربی و کُردی در یک متن طراحی می‌کنید. داده‌های مربوط به این تمرین در فایل ANN-HW1-Data.xlsx قرار گرفته است که حاوی ۵۰ جمله برای هر زبان است. از این داده، برای هر زبان، ۸۰٪ جملات اول (۴۰ جمله اول) را برای آموزش و مابقی ۲۰٪ (۱۰ جمله آخر) را برای آزمون جدا کنید.

الف) برای تشخیص این سه زبان، برای هر جمله تعداد چهار نویسه (کاراکتر) «ژ، پ، ل، ع» را به تعداد کل نویسه‌های آن جمله تقسیم کنید و بر اساس مقدار آن در مورد نوع زبان تصمیم بگیرید. برای این منظور، این بردار چهار بعدی را برای همه جملات آموزشی در هر زبان حساب کنید و میانگین این بردارهای آموزشی هر زبان به عنوان نماینده آن زبان استفاده کنید. بدین صورت که برای هر داده آزمون، هر زبانی که میانگین بردارهای نویسه آن بیشترین شباهت را با بردار جمله آزمون بر اساس معیار شباهت کسینوسی داشته باشد، به عنوان زبان آن جمله تشخیص داده می‌شود. معیار شباهت

h.veisi@ut.ac.ir

کسینوسی به صورت زیر است که در آن d1 وd2 دو بردار مورد مقایسه هستند. صورت کسر، ضرب داخلی دو بردار و مقادیر مخرج اندازه‌های دو بردار است.

$$\cos(\theta) = \frac{\mathbf{d}_1 . \mathbf{d}_2}{|\mathbf{d}_1| \, |\mathbf{d}_2|}$$

داده‌های مجموعه آزمون را برای ارزیابی روش خود به سیستم ارائه دهید و نتایج حاصل را با معیارهای صحت (Accuracy)، دقت (Precision)، یادآوری (Recall) و F-Measure گزارش کنید.

ب) می‌خواهیم کارایی سیستم را مقداری بهبود دهیم و یک بردار ویژگی بزرگ‌تر استفاده کنیم. به این منظور، از کل تعداد نویسه‌های متن به عنوان ویژگی استفاده می‌کنیم. به منظور طراحی این دسته‌بند ابتدا لازم است برداری با نام Character Frequency (CF) را معرفی کنیم. تعداد عناصر (مولفه‌های) این بردار برابر تعداد کل نویسه‌های موجود در تمامی سه زبان است. آنگاه برای هر جمله، عناصر این بردار برابر با فراوانی نرمال شده تعداد نویسه‌های آن متن استفاده می‌کنیم. به عنوان مثال فرض کنید نویسه‌های (A,B,C,D,E) تمامی نویسه‌های مجاز در سه زبان باشند. در این صورت، بردار CF متناظر با متن فرضی "ABBDCDEDEABB" به صورت (2, 4, 1, 3, 2) خواهد بود. حال با استفاده از بردار CF، بردار دیگری با نام Normal CF (NCF) با استفاده از فرمول $NCF(i) = CF(i)/\mathrm{Sum}(CF)$ تعریف می‌شود. بعد از نرمال کردن، بردار این جمله به صورت (0.17, 0.33, 0.08, 0.25, 0.17) خواهد بود (تقسیم مولفه‌ها بر ۱۲).

با داشتن این بردار برای هر جمله، میانگین همه بردارهای داده آموزش را برای هر زبان محاسبه کنید و از آن به عنوان معیار مقایسه داده‌های آزمون استفاده کنید. بدین صورت که برای هر داده آزمون، هر زبانی که میانگین بردارهای NCF آموزش آن، کم‌ترین فاصله کسینوسی را با NCF جمله آزمون داشته باشد، به عنوان زبان آن جمله تشخیص داده می‌شود.

معیارهای صحت (Accuracy)، دقت (Precision)، یادآوری (Recall) و F-Measure را در این حالت نیز

h.veisi@ut.ac.ir

به نام خدا

محاسبه کنید و با نتایج بخش الف مقایسه کنید و مشاهده و تحلیل خود را گزارش کنید.

ج) نتایج بخش ب را با ارزیابی مبتنی بر روش 5-fold Cross-Validation گزارش کنید. برای این کار کل مجموعه داده را استفاده کنید (و نه فقط مجموعه آموزش).