بسمه تعالی

تمرینات سری چهارم داده کاوی

۱. دو مجموعه داده CancerTraining و CancerTest ضمیمه شده‌اند. هر سطر در این فایل‌ها نشان‌دهنده اطلاعات یک بیمار است که با کاما از هم جدا شده‌اند. اولین عدد، id بیمار است و آخرین مقدار تشخیص خوش خیم (benign) یا بدخیم بودن (malignant) تومور است. سایر خصیصه ها مقادیر بین ۱ تا ۱۰ دارند و به ترتیب عبارتند از:

Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion,
Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses

چندین روش برای ساخت درخت تصمیم (مثل ID3) و همین طور روش های مختلف برای هرس درخت را بکار ببرید و جواب ها را با هم مقایسه کنید.

۲. در دسته بندی متن هدف این است که موضوع یک متن (خبر، مقاله، وبلاگ، ...) مشخص شود. سه فایل training, test, topics ضمیمه شده اند. در هر یک از دو فایل training, test متن‌هایی به فرمت زیر قرار دارند: (خط اول موضوع، سپس یک خط خالی، عنوان، خط خالی، محل و تاریخ، خط خالی، متن اصلی)

topic (classification)
blank line
title
blank line
location, date
blank line
text

موضوعات مختلفی که متن‌ها می توانند داشته باشند در فایل topics قرار دارد. هدف این است که موضوع هر یک از متن های داخل فایل test را با استفاده از الگوریتم k-nearest neighbor پیش بینی کنید. از توابع فاصله یا شباهت زیر و نمایش‌های نظیر آن‌ها استفاده کنید.

(a) فاصله Hamming: هر متن را با یک بردار دودویی نمایش دهید که هر بیت نشان دهنده این است که آیا کلمه نظیر در متن ظاهر شده یا خیر.

(b) فاصله اقلیدسی: هر متن با یک بردار عددی نمایش داده می‌شود که هر عدد نشان دهنده این است که کلمه نظیر چند بار در متن ظاهر شده (می تواند صفر باشد).

(c) شباهت کسینوسی: بردارهای عددی TF-IDF به روشی که توضیح داده می‌شود برای متن‌ها ساخته می‌شوند. سپس شباهت بین دو بردار بصورت کسینوس زاویه بین آن‌ها (ضرب داخلی دو بردار تقسیم بر حاصلضرب نرم آن‌ها) تعریف می‌شود.

فرض کنید w یک کلمه و d یک متن باشد؛ N(d,w) تعداد رخداد کلمه w در متن d باشد؛ W(d) تعداد کل کلمات متن d باشد؛ TF مخفف term frequency است و عبارتست از TF = N(d,w)/W(d).

فرض کنید D تعداد کل متن‌ها باشد و C(w) تعداد متن‌هایی باشد که شامل کلمه w هستند. IDF مخفف Inverted document frequency است و عبارتست از IDF(d,w) = log(D/C(w)) . وزن TF-IDF برای کلمه w در متن d بصورت مقابل تعریف می‌شود: TF(d,w)*IDF(d,w). این عددی است که در موقعیت هر کلمه w در بردار نظیر متن d باید بگذارید. (این یک معیار هیوریستیک است که سعی دارد محتوای اطلاعات هر کلمه را مشخص کند. بنابراین کلمه ای مثل the که در تمام متن ها ظاهر می-شود IDF برابر با صفر دارد که نرخ زیاد حضور آن در متن را خنثی می‌کند. از طرف دیگر کلماتی که خاص همان متن هستند، تقویت می‌شوند.

برای هر یک از سه نمایش بالا k=1, 3, 5 را امتحان کنید. دقت کنید نزدیک ترین همسایه کسی است که فاصله اش کمتر و یا شباهتش بیشتر باشد.

گزارش خود از نتایج تمرین را با جداول و نمودارهایی نمایش دهید.

- حل تمرینات خود را در یک فایل فشرده به نام DM_Assignment04_names بریزید و به آدرس taherian.khu@gmail.com ارسال کنید. عنوان ایمیل را DM_Assignment04_names بگذارید. (توجه کنید باید به جای names، اسامی افراد تیم را بگذارید.) دقت کنید عنوان را فراموش نکنید وگرنه ایمیل بررسی نمی شود.
- مهلت ارسال این تمرینات تا تاریخ سه شنبه ۲۶ اردیبهشت ۱۳۹۶ قبل از ساعت ۸ صبح می باشد. بعد از این تاریخ به هیچ عنوان پذیرفته نمی شود و نمره ای تعلق نمی گیرد. بنابراین سعی کنید تمرینات را تا یکی دو روز قبل از اتمام مهلت انجام دهید تا با مشکلاتی مثل قطعی اینترنت مواجه نشوید. مشکلات این چنینی به هیچ عنوان پذیرفته نیست.