دانشکده‌ی مهندسی کامپیوتر

# آمار و احتمال مهندسی

**نیم‌سال اول ۱۳۹۹ ـ ۱۴۰۰**

مدرس: امیر نجفی

تمرین عملی

## ۱ پیش‌گفتار

در این تمرین قصد داریم شما را با روند شبیه‌سازی مسائل در آمار و احتمالات آشنا کرده و کاربرد این مباحث در برخی کارهای روزمره را نیز به شما نشان دهیم. از آنجایی که این تمرین نیاز به پیاده‌سازی دارد، لطفا به نکات زیر توجه فرمایید:

- برای پیاده‌سازی پاسخ کلیه مسائل می‌توانید از زبان‌های R ، Python و MATLAB استفاده کنید. البته استفاده از زبان R ، ۰٫۲۵ نمره امتیازی دارد.
- برای پیاده‌سازی می‌توانید یکی از سوالات زیر را به دلخواه انتخاب کنید و ارسال پاسخ همان یک سوال کافی است. در صورتی که مسئله «تشخیص ایمیل‌های سالم از هرزنامه» را انتخاب کرده و آن را به درستی و با کیفیت بالایی انجام بدهید، می‌توانید تاحداکثر ۰٫۵ نمره اضافی نیز از آن بابت دریافت نمایید.
- ارسال کد پیاده‌سازی به تنهایی کفایت نمی‌کند. لطفا به پیوست کد، گزارشی حداقل ۵ صفحه‌ای از روند پیاده‌سازی کار و نتایج خود را ارسال نمایید. اگر در سوالی نموداری خواسته شده و یا برخی معیارهای کمی برای ارزیابی سیستم‌ها پیشنهاد شده است، آن‌ها را حتما در گزارش بیاورید.
- لطفا کدها را به نحوی بنویسید که خوانایی خوبی داشته باشند. استفاده از کامنت‌ها و نام‌های معقول برای متغیرها توصیه می‌گردد.
- مهلت ارسال پاسخ این تمرین، تا پایان روز ۲۰ام فروردین ماه ۱۴۰۰ می‌باشد (اکیداً غیرقابل تمدید).

## ۲ مسائل

### مسئله‌ی ۱. مشاهده یک فرآیند احتمالی

هدف از این تمرین مشاهده یک فرآیند احتمالی است. فرض کنید یک مربع واحد دارید و بطور تصادفی نقطه‌ای درون آن قرار می‌گیرد. سپس با سیاستی که در ادامه می‌آید مکان آن را تغییر می‌دهیم. فاصله این نقطه تا چهار ضلع مربع اولیه را در نظر می‌گیریم و کوتاه‌ترین فاصله را حساب می‌کنیم. سپس مربعی به اندازه $0 < p < 1$ برابر این فاصله حول مربع حساب کرده و نقطه‌ای تصادفی بصورت یکنواخت داخل آن انتخاب می‌کنیم.

**نکته:** هر بار فاصله نقطه تا مربع اولیه حساب می‌شود.

با توجه به مسئله به سوالات زیر پاسخ دهید:

الف: فرض کنید $p = \frac{1}{2}$، آنگاه به صورت تصویری (ترجیحاً از طریق یک نقشه گرمایی یا heatmap) احتمال رخ‌دادن هر نقطه در مربع در طی فرآیند فوق را مشخص کنید. برای این کار می‌توانید مربع را به زیرمربع‌های ریز و بدون همپوشانی تقسیم کرده و تعداد دفعاتی که نقطه تصادفی در هر کدام واقع می‌شود را بشمارید و در آخر به احتمال تبدیل کنید. همچنین، احتمالا لازم خواهد بود که شبیه‌سازی را چندین بار تکرار کنید تا نقشه گرمایی خوبی بدست بیاورید.

ب: این مسئله اثبات نظری برای همگرایی دارد. در یک اجرا مسیر حرکت نقطه را در داخل مربع نشان دهید. همگرایی در کجای مربع اتفاق می‌افتد؟

پ: نمودار زمان همگرایی بر حسب $p$ های مختلف را رسم کنید. برای محاسبه زمان همگرایی هر $p$ بهتر است برای هر مقدار ۱۰۰ بار اجرا انجام داده سپس میانگین بگیرید. فاصله مقادیر متوالی $p$ نیز ۰/۰۵ در نظر گرفته شود.

## مسئله‌ی ۲. تشخیص ایمیل‌های سالم از هرزنامه[1]

### توضیحات اولیه

هرزنامه‌ها شامل ایمیل‌های تبلیغاتی، یا متونی با اهداف خرابکارانه هستند که معمولا کلمات خاصی را شامل می‌شوند. مثلاً ممکن است در چنین ایمیل‌هایی به شما گفته شود که قوانینی را نقض کرده‌اید و برای اینکه جریمه نشوید باید سریعاً روی یک لینک کلیک کنید. کلماتی مانند «لینک» را در این ایمیل‌ها زیاد می‌بینید. احتمال مشاهده برخی کلمات خاص، به ما اطلاعات خوبی برای پیشبرد روند استنتاج می‌دهد؛ به گونه‌ای که اگر برخی کلمات را در یک متن بیشتر مشاهده کنیم، می‌توانیم با احتمال قابل توجهی آن متن را هرزنامه در نظر بگیریم. در مقابل، عدم رخداد آن کلمات، نشان از یک ایمیل عادی دارد. در این تمرین قصد داریم سیستمی بر پایه قانون بیز طراحی کنیم که بتواند ایمیل‌های سالم را از آن‌هایی که هرزنامه هستند، تشخیص دهد.

### داده‌های مورد استفاده

شما از مجموعه ۶۰۰ ایمیل برای طراحی (یادگیری) این سیستم و ۴۰۰ ایمیل برای ارزیابی آن استفاده خواهید کرد. کلیه ایمیل‌ها در قالب فایل‌های متنی در لینک زیر موجود هستند:

**لینک مشاهده مجموعه ایمیل‌ها**

در این صفحه ۴ پوشه با نام‌های hamtraining و spamtraining برای طراحی و hamtesting و spamtesting برای ارزیابی سیستم موجود هستند. همانطور که از نام پوشه‌ها مشخص است، پوشه‌ای که با کلمه ham مشخص می‌شود، شامل تعدادی ایمیل عادی و پوشه‌ای که نامش با spam آغاز می‌گردد، شامل تعدادی هرزنامه است.

---

[1] spam

## مکانیزم محاسبه احتمالات

فرض کنید $E$ یک ایمیل شامل کلمات $w_1 w_2 \dots w_n$ باشد و رخداد یک ایمیل عادی و یک هرزنامه به ترتیب با حروف $H$ و $S$ نمایش داده شود. با این نمادگذاری، احتمال اینکه $E$ یک ایمیل عادی باشد، $P(H|E)$ و احتمال هرزنامه‌بودن آن برابر $P(S|E)$ است. احتمال رخداد ایمیل $E$ نیز با $P(E)$ مشخص می‌شود. برای سادگی فرض کنید داریم:

$$P(E) = P(w_1 w_2 \dots w_n) = \prod_{i=1}^{n} P(w_i)$$

(توضیح دهید چرا این فرض در دنیای واقعی اشتباه است). حال برای محاسبه احتمال عادی‌بودن یا هرزنامه‌بودن $E$ می‌توانیم از احتمال مشاهده کلمات آن در ایمیل‌های عادی و هرزنامه بهره بگیریم. به طور دقیق‌تر می‌توانیم بنویسیم:

$$P(E|S) = P(w_1 w_2 \dots w_n|S) = \prod_{i=1}^{n} P(w_i|S)$$

$$P(E|H) = P(w_1 w_2 \dots w_n|H) = \prod_{i=1}^{n} P(w_i|H)$$

اگر $w$ یک کلمه باشد، برای محاسبه احتمال $P(w|H)$ و $P(w|S)$ نیز می‌توانیم از فرمول محاسبه احتمال شرطی به کمک اشتراک بهره بگیریم:

$$P(w|H) = \frac{P(w \cap H)}{P(H)}$$

$$P(w|S) = \frac{P(w \cap S)}{P(S)}$$

نهایتا ما به دنبال $P(S|E)$ و $P(H|E)$ هستیم که به طریق زیر محاسبه می‌گردند:

$$P(S|E) = \frac{P(E|S)P(S)}{P(E)}$$

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

اگر $P(S|E)$ بزرگتر از $P(H|E)$ باشد، گوییم ایمیل هرزنامه است و در غیر این صورت آن را سالم می‌دانیم.

برای محاسبه این احتمالات ابتدا کلیه ایمیل‌های دو پوشه hamtraining و spamtraining را برای طراحی سیستم به کار بگیرید. یعنی فرض کنید کلیه دانش قبلی شما درباره ایمیل‌ها محدود به این دو پوشه است. حال یک ایمیل از یکی از پوشه‌های hamtesting یا spamtesting به شما داده می‌شود؛ به طوری که نمی‌دانید این ایمیل از کدام پوشه برداشته شده است. طبق مکانیزمی که گفته شد، برای این ایمیل که فقط متن آن در اختیار سیستم شما قرار می‌گیرد و سیستم از عادی یا هرزنامه‌بودن آن نامطلع است، احتمالات عادی‌بودن یا هرزنامه‌بودن را حساب کنید و نهایتا هر ایمیل موجود در این پوشه‌های testing را دسته‌بندی نمایید؛ مثلا یک لیست متناظر ایمیل‌های ۱ تا ۲۰۰ از پوشه hamtesting درست کنید که $i$ ـامین عنصر آن، نشان بدهد ایمیل $i$ ـام از این پوشه توسط سیستم به عنوان ایمیل عادی شناخته شده یا در دسته هرزنامه قرار گرفته است. همین کار را برای پوشه spamtesting نیز

انجام بدهید. نهایتا شما متناظر هر کدام از ۴۰۰ ایمیل موجود در پوشه‌های hamtesting و spamtesting یک دسته‌بندی خواهید داشت. از طرفی، برای هر ایمیل دسته‌بندی صحیح آن (عادی یا هرزنامه) نیز مشخص است. لذا می‌توانید خروجی سیستم را با واقعیت روی زمین مقایسه کنید.

بدین منظور تعداد ایمیل‌هایی را که سیستم شما به درستی دسته‌بندی کرده، بر تعداد کل ایمیل‌ها تقسیم کرده، نتیجه را در پیاده‌سازی آورده و در گزارش خود ذکر کنید.

## برخی ملاحظات برای پیاده‌سازی

در حین پیاده‌سازی خواسته‌های این سوال، لطفا رخداد هر کاراکتر فاصله را به عنوان مشخص‌کننده اتمام هر کلمه در نظر بگیرید. می‌توانید برخی کاراکترها مانند نقطه ویرگول و ... را نیز حذف کنید و فقط کلمات با طول حداقل ۲ کاراکتر را نگه داشته و بقیه را کنار بگذارید. ممکن است کلمات بسیاری در ایمیل‌ها وجود داشته باشند که بعضی از آن‌ها تعداد رخداد بسیار کمی دارند. لذا می‌توانید تنها ۵۰۰ کلمه با بیشترین تکرار را در نظر بگیرید و رخداد بقیه کلمات برایتان بی‌اهمیت باشد.

توجه کنید که استفاده از کتابخانه‌های آماده برای محاسبه احتمالات مذکور، در این سوال مجاز نیست. اما می‌توانید از کتابخانه‌های دیگر که برای مقاصد به جز این استفاده می‌گردند، بهره بگیرید.