



(برای یک جلسه از درس، برای مثال ۲ ساعت از کلاس درس در یک هفته)

دانشکده:	بهداشت	گروه آموزشی:	آمار زیستی	مقطع و رشته‌ی تحصیلی:	کارشناسی ارشد
نام درس:	اصول و روش‌های داده کاوی	تعداد واحد: ۲	نوع واحد: ۲ واحد نظری - ۰ واحد عملی	پیش نیاز:	
زمان برگزاری کلاس:	یکشنبه	ساعت: ۱۰-۱۳	مکان برگزاری: -		
تعداد دانشجویان:	۲	مسئول درس:	مدرسین (به ترتیب حروف الفبا): دکتر مسعود		
صالحی - مرتضی محمدزاده					

شرح درس: (لطفاً شرح دهید)

این درس مقدمه‌ای جامع بر روش‌های نوین یادگیری آماری (آمار محاسباتی و یادگیری ماشین) با تأکید بر کاربردهای آن در داده‌های زیستی و پزشکی است. با توجه به رشد انفجاری داده‌های با ابعاد بالا (مانند داده‌های ژنومی و تصویربرداری پزشکی)، روش‌های کلاسیک آماری پاسخگوی نیازها نیستند. این درس به بررسی روش‌های یادگیری با نظارت (Supervised) و بدون نظارت (Unsupervised)، انتخاب مدل، تنظیم‌سازی (Regularization)، روش‌های بازنمونه‌گیری (Resampling) و تحلیل بقا می‌پردازد. تمامی مباحث با انجام عملی در محیط نرم‌افزار R و بر روی داده‌های واقعی زیستی همراه خواهد بود.

هدف کلی: (همان هدف بینابینی طرح دوره است)

هدف کلی این درس، توانمندسازی دانشجویان کارشناسی ارشد آمار زیستی در تحلیل داده‌های پیچیده و حجیم زیستی، انتخاب مدل‌های پیش‌بینی مناسب، ارزیابی دقت مدل‌ها و تفسیر صحیح نتایج در بافت بالینی و اپیدمیولوژیک است.



در پایان دوره انتظار می‌رود که دانشجو موارد ذیل را فراگیرد:

۱. درک عمیق مصالحه بایاس-واریانس (Bias-Variance Trade-off) و ارزیابی دقت مدل‌ها.
۲. تسلط بر روش‌های بازنمونه‌گیری Cross-Validation و Bootstrap برای انتخاب مدل و برآورد خطا.
۳. توانایی به‌کارگیری روش‌های انتخاب متغیر و منظم‌سازی (Ridge, Lasso, PCR) در داده‌های با ابعاد بالا ($p > n$).
۴. پیاده‌سازی و تفسیر مدل‌های طبقه‌بندی (Logistic, LDA, QDA, Naive Bayes, SVM) در داده‌های بالینی.
۵. تحلیل داده‌های بقا و داده‌های سانسور شده با استفاده از مدل مخاطرات متناسب کاکس و روش‌های درختی.
۶. انجام تحلیل‌های بدون ناظر (PCA) و خوشه‌بندی (برای کشف الگوها در داده‌ها).
۷. پیاده‌سازی تمامی روش‌های فوق با کدنویسی در نرم‌افزار R.

شیوه‌های تدریس:

- سخنرانی و ارائه مطالب نظری
- حل مسائل و مثال‌های واقعی زیستی
- کارگاه‌های عملی و کدنویسی در محیط R (Computer Labs)
- بحث گروهی و بررسی مقالات مرتبط (Journal Club)

وظایف و تکالیف دانشجو: (لطفاً شرح دهید)

- حضور فعال در کلاس‌های نظری و عملی
- انجام تمرینات کدنویسی در R پس از هر جلسه
- تحلیل یک مجموعه داده واقعی زیستی به عنوان پروژه پایانی
- شرکت در امتحان میان‌ترم و پایان‌ترم

وسایل کمک آموزشی:

- ویدئو پروژکتور و وایت‌برد
- کامپیوترهای مجهز به نرم‌افزار R و RStudio
- اسلایدهای آموزشی و جزوات

نحوه ارزشیابی و درصد نمره: (از نمره کل)

- آزمون میان‌ترم (۳۰٪) تشریحی و محاسباتی
- آزمون پایان‌ترم (۴۰٪) تشریحی و محاسباتی
- تمرینات و پروژه عملی (R: ۳۰٪) شامل تحلیل داده‌های واقعی
- نوع آزمون: تشریحی، چند گزینه‌ای و پروژه‌محور

به نام خداوند جان آفرین



دانشگاه علوم پزشکی ایران
مرکز مطالعات و توسعه آموزش علوم پزشکی
واحد برنامه‌ریزی آموزشی
طرح درس (Lesson Plan)

منابع پیشنهادی برای مطالعه: (لطفا نام ببرید):

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R* (2nd ed.). Springer.
2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer.



جدول هفتگی کلیات ارائه ی درس

هفته	عنوان مطلب	استاد / اساتید	فعالیت‌ها
۱	مقدمه‌ای بر یادگیری آماری، تفاوت یادگیری با نظارت و بدون نظارت، KNN	دکتر مرتضی محمدزاده	آشنایی با محیط R و RStudio، بارگذاری داده‌های زیستی (Lab 2)
۲	مرور رگرسیون خطی، ارزیابی دقت مدل (RSE, R^2) و تشخیص مشکلات مدل	دکتر مرتضی محمدزاده	پیاپی سازی رگرسیون خطی بر روی داده‌های بالینی (Lab 3)
۳	طبقه‌بندی: رگرسیون لجستیک، مدل‌های خطی تعمیم یافته (GLM)	دکتر مرتضی محمدزاده	تحلیل داده‌های دوحالتی (مثلاً ابتلا به بیماری) در R (Lab 4)
۴	طبقه‌بندی: Naive Bayes, QDA, LDA و KNN	دکتر مرتضی محمدزاده	مقایسه روش‌های طبقه‌بندی بر روی داده‌های میکرواری
۵	حل تمرین و رفع اشکال (فصل ۲، ۳ و ۴)	دکتر مرتضی محمدزاده	مرور مباحث نظری، حل مسائل تحلیلی و رفع اشکال کدهای R دانشجویان
۶	روش‌های باز نمونه‌گیری: Cross-Validation و Bootstrap	دکتر مرتضی محمدزاده	شبیه‌سازی و ارزیابی دقت مدل‌ها با تکنیک‌های Resampling (Lab 5)
۷	انتخاب مدل خطی و تنظیم‌سازی: Best Subset, Ridge, Lasso	دکتر مرتضی محمدزاده	تحلیل داده‌های ژنومی با ابعاد بالا ($n < p$) در R (Lab 6)
۸	ادامه تنظیم‌سازی: PCR و PLS + حل تمرین و رفع اشکال (فصل ۵ و ۶)	دکتر مرتضی محمدزاده	مقایسه Ridge, Lasso و PCR در داده‌های واقعی زیستی
۹	آزمون میان ترم (پوشش مباحث هفته‌های ۱ تا ۸)	دکتر مرتضی محمدزاده	ارزیابی کتبی و محاسباتی
۱۰	فراتر از خطی بودن: رگرسیون چند جمله‌ای، Step Functions و Splines	دکتر مرتضی محمدزاده	مدل سازی غیرخطی داده‌های بالینی (Lab 7)
۱۱	مدل‌های جمعی تعمیم یافته (GAMs) و کاربرد آن در اپیدمیولوژی	دکتر مرتضی محمدزاده	پیاپی سازی GAM و تفسیر نمودارهای آن در R
۱۲	روش‌های درختی: درخت‌های تصمیم، Bagging و Random Forest	دکتر مرتضی محمدزاده	تحلیل داده‌های بقا و طبقه‌بندی با درخت‌ها (Lab 8)
۱۳	Boosting, BART و ماشین‌های بردار پشتیبان (SVM)	دکتر مرتضی محمدزاده	پیاپی سازی SVM و Boosting در داده‌های پزشکی (Lab 9)
۱۴	تحلیل بقا و داده‌های سانسور شده: منحنی کاپلان-مایر و مدل کاکس	دکتر مرتضی محمدزاده	تحلیل داده‌های کارآزمایی بالینی و بقای بیماران (Lab 11)

به نام خداوند جان آفرین



دانشگاه علوم پزشکی ایران
مرکز مطالعات و توسعه آموزش علوم پزشکی
واحد برنامه‌ریزی آموزشی
طرح درس (Lesson Plan)

۱۵	یادگیری بدون نظارت: PCA، خوشه‌بندی و آزمون‌های چندگانه	دکتر مرتضی محمدزاده	تحلیل داده‌های NCI60 و آشنایی با FDR (Lab 12 & 13)
۱۶	حل تمرین و رفع اشکال نهایی و ارائه پروژه‌ها	دکتر مرتضی محمدزاده	مرور کل مباحث، رفع اشکال کدهای پروژه پایانی و آمادگی برای امتحان پایان‌ترم