

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



دانشگاه فردوسی مشهد
دانشکده علوم
گروه آمار

صورتجلسه دفاع رساله کارشناسی ارشد آمار ریاضی

در تاریخ ۲۵/۴/۲۸ خانم / آقای سیدرضا کاوشی از رساله کارشناسی ارشد خود
تحت عنوان :

"تحلیل ممیزی و کاربرد آن"

با بیان خلاصه ای از کار انجام شده و پاسخ به سئوالات داوران دفاع نمودند و

این رساله با نمره ۲۵/- معادل عالی قبول شد.

- ۱- استادراهنما دکتر حسن صادقی
- ۲- اعضاء هیئت داوران
 - ۱- دکتر علی عمیدی (داور مدعو)
 - ۲- دکتر ناصر رضا ارقامی (استاد مشاور)
- ۳-

معاون آموزشی دانشکده

مدیر گروه آمار

دکتر غلامحسین شاهکار



رساله کارشناسی ارشد

آمار ریاضی

تحلیل همبستگی و کاربرد آن

استاد راهنما:

دکتر حسن صادقی (دانشیار گروه آمار دانشگاه فردوسی)

استاد مشاور:

دکتر ناصر رضا ارقامی (دانشیار گروه آمار دانشگاه فردوسی)

استاد داور:

دکتر علی عمیدی (استاد گروه آمار دانشگاه شهید بهشتی)

نگارش:

سیدرضا کاوشی

بهار ۱۳۷۵

3345/۲

تقدیم به کسانی که لذت زندگی را در استفاده از علم می جویند.

تقدیم به آنان که برای به کارگیری علم آمار، همچنان می کوشند
و از جان مایه می گذارند.

تقدیم به تمامی عزیزانم...

تشکر و قدردانی

در ابتدا، بر خود، لازم می‌دانم از پدر و مادر بزرگواری که اینجانب را با صبر و دلسوزی فراوان مورد توجه قرار داده‌اند؛ و از تمامی اساتید گروه آمار، به خصوص از استاد ارجمند و مهربانم، جناب آقای دکتر حسن صادقی (دانشیار گروه آمار دانشگاه فردوسی) به عنوان یک معلم و استاد تمام عیار و بواسطه راهنمایی‌های ارزنده ایشان در تمامی طول رساله؛ و از مشاورت استاد محترم جناب آقای دکتر ارقامی (دانشیار گروه آمار دانشگاه فردوسی - استاد مشاور)؛ و همچنین از استاد عالی‌مقام جناب آقای پروفسور دکتر علی عمیدی (استاد دانشگاه شهید بهشتی و ویراستار مرکز نشر دانشگاهی) که دعوت و داوری این رساله را پذیرفتند و بواسطه اشاره‌های ذی‌قیمت ایشان به نقاط ضعف این پایان‌نامه، نهایت تشکر و قدردانی را بنمایم.

همچنین، از آقای دکتر درویش (استاد بخش بیوسیستماتیک گروه زیست‌شناسی دانشکده علوم دانشگاه فردوسی مشهد) که داده‌های تحقیق خود را برای انجام تحلیلهای متعدد ممیزی در اختیار اینجانب قرار دادند؛ و از آقایان پروفسور، دکتر سیروس رضانی (استاد دانشگاه ویسکانسین مدیسن امریکا) و دکتر غلامرضا نخعی‌زاده (استاد دانشگاه و محقق ارشد شرکت دایملر بنز آلمان) که به ترتیب در جهت تهیه و ارسال برنامه‌های آماری فصل پنجم (FACT, S-PLUS) و C4.5، زحمات بی‌شائبه‌ای را متقبل گردیدند؛ و همچنین از زحمات و مساعدت‌های دوستان عزیزم آقایان رسول اتحاد (مسئول محترم کتابخانه دانشکده علوم ریاضی دانشگاه مشهد)، علیرضا وطن‌دوست (مسئول آزمایشگاه آمار دانشکده علوم ریاضی)، و مجید سرمد (مربی محترم گروه آمار دانشگاه فردوسی) و از خانم و آقای پاکرو (معاونت پژوهشی دانشگاه فردوسی) بیشترین مراتب سپاس و امتنان خود را اعلام دارم.

سیدرضا کاوشی

بهار - ۱۳۷۵

پیشگفتار

با توجه به اهمیت و ضرورت بکارگیری روشهای تحلیل ممیزی آماری، در زمینه‌های گوناگون تحقیقاتی، تدوین و ارائه منبعی که بتواند نیازهای استفاده‌کنندگان را برآورده نماید، ضروری به نظر می‌رسد. این خلاصه که در جهت رفع نیاز مذکور تدوین شده است، به طور رسمی دارای پنج، و به طور غیر رسمی دارای هفت فصل می‌باشد. در تمامی طول رساله سعی شده است که موضوعات مطروحه به ساده‌ترین شیوه بیان گردیده تا استفاده‌کنندگان محترم بتوانند نیازهای خود را رفع نمایند. برای این منظور در پایان هر فصل نتایج عملی ضمیمه گردیده است.

در فصل اول، و ما به معرفی و بررسی تحلیل ممیزی خطی فیشر پرداخته‌ایم. در این فصل عمدتاً به مفهوم فاصله‌های آماری توجه شده است.

در فصل دوم، مقدمه‌ای را برای ممیزی، با استفاده از اطلاعات دیگری همچون توزیعهای احتمالی بردار اندازه، آغاز کرده و برخی از حالت‌های خاص را مورد توجه قرار داده‌ایم. در این فصل یک روش پیشنهادی را برای برآورد احتمالات رده‌بندی نادرست روش فیشر در دو جامعه ارائه نموده و تعمیم آن را در فصل ششم آورده‌ایم.

در فصل سوم، ابتدا مطالب فصل دوم را به حالت بیشتر از دو جامعه تعمیم داده و روش فیشر را به روش بردارهای متعارف بسط و خواص آن را مورد بررسی قرار داده‌ایم. در ادامه، روشهای ممیزی لجستیک و دنباله‌ای آمده است.

به طور کلی مطالب سه فصل اول، کلیات و اصول اولیه ممیزی را شامل می‌شوند، که در فصول بعد مورد استفاده قرار خواهند گرفت.

در فصل چهارم، مقدمه‌ای بر روشهای تحلیل ممیزی اطلاعات کیفی را آغاز نموده و در فصل پنجم مباحث را با درختهای کلاس‌بندی ادامه داده‌ایم.

لازم به توضیح است، مواردی که با (*) مشخص شده است، بوسیله اینجانب و به راهنمایی استاد گرامیم جناب آقای دکتر صادقی، بیان و یا اثبات گردیده است. و مواردی که با (***) مشخص گردیده، کاملاً نو هستند.

فهرست مطالب

صفحه	عنوان
	فصل اول : تابع ممیز خطی
۱-۱	۱-۱- مقدمه
۱-۱	۲-۱- تعریف تحلیل ممیزی
۲-۱	۱-۲-۱- مراحل تحلیل ممیزی
۳-۱	۲-۲-۱- ارائه چند زمینه کاربردی
۵-۱	۳-۱- روشهای غیر رسمی ممیزی
۶-۱	۴-۱- نظریه تابع ممیزی خطی
۱۰-۱	۱-۴-۱- دستورات محاسبه احتمالات خطا
۱۳-۱	۲-۴-۱- ارائه دو مثال
۱۶-۱	۳-۴-۱- توابع ممیز خطی فیشر
۱۷-۱	۴-۴-۱- معایب و مزایای تابع ممیز خطی
۱۸-۱	۵-۱- وجه تشابه تابع ممیز خطی و الگوهای خطی
۲۱-۱	۱-۵-۱- ارتباط تحلیل ممیزی و مسأله مقایسه میانگینهای دو جامعه
۲۳-۱	۲-۵-۱- ارتباط D^2 و R^2
۲۳-۱	۳-۵-۱- چند تذکر
۲۷-۱	۶-۱- انتخاب متغیرها
۲۸-۱	۱-۶-۱- آزمونهای کفایت زیر مجموعه‌ای از متغیرها

- ۳۰-۱ - ۲.۶.۱- انتخاب متغیرها با شیوه مینیم کردن لامبدای وبلکس
- ۳۲-۱ - ۳.۶.۱- نحوه تفسیر ضرایب همبستگی در مفید بودن یک متغیر
- ۳۴-۱ - ۴.۶.۱- راه حل نظری انتخاب کاراترین متغیرها
- ۳۵-۱ - ۱.۴.۶.۱- اثبات الف به ب
- ۳۶-۱ - ۲.۴.۶.۱- اثبات ب به الف
- ۳۷-۱ - ۳.۴.۶.۱- اثبات ب به پ
- ۳۷-۱ - ۴.۴.۶.۱- اثبات پ به الف
- ۳۸-۱ - ۵.۴.۶.۱- اثبات الف به ت
- ۳۸-۱ - ۶.۴.۶.۱- اثبات ت به ب
- ۳۹-۱ - ۷.۴.۶.۱- اثبات ث به الف
- ۳۹-۱ - ۸.۴.۶.۱- اثبات الف به ث
- ۴۰-۱ - ۵.۶.۱- سایر روشهای انتخاب متغیرها
- ۴۱-۱ - ۷.۱- خروجیهای کامپیوتری

منابع فصل اول

فصل دوم: بهینه‌سازی تابع ممیز

- ۱-۲ - ۱.۲- روش مینیم کردن احتمال کل رده‌بندی نادرست
- ۳-۲ - ۲.۲- ممیزی به روش مینیماکس
- ۵-۲ - ۳.۲- ارتباط میزان خطای بهینه برای توزیع نرمال با P_1 و δ^2
- ۶-۲ - ۴.۲- قاعده ممیزی وقتی اندازه یکی از احتمالات خطا از پیش تعیین می‌شود
- ۷-۲ - ۵.۲- ممیزی به روش بیز
- ۸-۲ - ۱.۵- صورت محاسباتی دستور بیز
- ۹-۲ - ۶.۲- ممیزی با معیار مینیم کردن امید ریاضی یا متوسط زیان رده‌بندی نادرست
- ۱۰-۲ - ۷.۲- قاعده ممیزی مینیماکس با در نظر گرفتن زیانهای رده‌بندی نادرست

- ۱۱-۲ - ۸.۲- معیارهای برآورد خطا
- ۱۱-۲ - ۱.۸.۲- معیارهای نظری برآورد خطا
- ۱۳-۲ - ۲.۸.۲- معیارهای عملی برآورد خطا
- ۱۷-۲ - ۹.۲- یک روش پیشنهادی برای برآورد احتمالات رده بندی نادرست
- ۱۷-۲ - ۱۰.۲- استفاده از معیارهای برآورد خطا در تعدیل میزانهای خطا
- ۱۸-۲ - ۱.۱۰.۲- بهینه سازی تابع ممیز خطی در دو جامعه
- ۱۹-۲ - ۱۱.۲- یک روش ناپارامتری برای انتخاب متغیرها
- ۲۰-۲ - ۱۲.۲- روش محک اطلاع آکائیک و مخاطره طبیعی برای انتخاب متغیرها
- ۲۳-۲ - ۱۳.۲- ممیزی درجه دوم
- ۲۵-۲ - ۱.۱۳.۲- چند حالت، خرابی
- ۲۷-۲ - ۲.۱۳.۲- محاسن و معایب تابع ممیز درجه دوم
- ۲۸-۲ - ۱۴.۲- بهترین تابع ممیز خطی
- ۳۱-۲ - ۱۵.۲- خروجیهای کامپیوتری
- ۴۷-۲ - منابع فصل دوم

فصل سوم: ممیزی در G جامعه و سایر روشهای ممیزی

- ۱-۳ - ۱.۳- قاعده ممیزی بهینه در G جامعه
- ۳-۳ - ۲.۳- قاعده ممیزی بیز در G جامعه
- ۴-۳ - ۳.۳- قاعده ممیزی مینیماکس در G جامعه
- ۵-۳ - ۴.۳- شکل دیگر قاعده ممیزی بهینه
- ۵-۳ - ۵.۳- قاعده بهینه با در نظر گرفتن زبانهای رده بندی نادرست
- ۹-۳ - ۶.۳- ممیزی با استفاده از بردارهای متعارف و قضایای مربوط به آن
- ۱۹-۳ - ۷.۳- تحلیل ممیزی لجستیک
- ۱۹-۳ - ۱.۷.۳- الگوی رگرسیون لجستیک

۲۰-۳	۲-۷-۳- تفسیر ضرایب
۲۰-۳	۳-۷-۳- برآورد پارامترها
۲۲-۳	۱-۳-۷-۳- وجود برآوردگرهای درست‌نمایی ماکزیم
۲۵-۳	۴-۷-۳- گسترش توابع لجستیک چند متغیره
۲۵-۳	۱-۴-۷-۳- حالت خاص
۲۶-۳	۲-۴-۷-۳- بهینه‌سازی الگوی لجستیک
۲۷-۳	۵-۷-۳- تعمیم الگوی لجستیک به G جامعه
۲۷-۳	۶-۷-۳- نقاط ضعف وقوت
۲۸-۳	۷-۷-۳- مقایسه تابع ممیز خطی و الگوی لجستیک در مثال فیشر
۲۸-۳	۸-۳- تحلیل ممیزی دنباله‌ای
۳۰-۴	۱-۸-۳- روش اول
۳۱-۳	۲-۸-۳- روش دوم
۳۳-۳	۳-۸-۳- روش سوم
۳۳-۳	۴-۸-۳- روش چهارم
۳۴-۳	۹-۳- سایر روشها
۳۵-۳	۱۰-۳- خروجیهای کامپیوتری
۱۰۳-۳	منابع فصل سوم

فصل چهارم: ممیزی داده‌های کیفی

۱-۴	۱-۴- مقدمه
۱-۴	۲-۴- متغیرهای کیفی
۳-۴	۱-۲-۴- تبدیل یک متغیر تصادفی گسسته به متغیرهای برنولی
۴-۴	۳-۴- استفاده از LDF برای ممیزی گسسته
۴-۴	۱-۳-۴- مثال نقض

۴-۴	الگوهای چند جمله‌ای کامل
۶-۴	۱-۴-۴- تعمیم به حالت G جامعه
۶-۴	۲-۴-۴- نمونه‌گیری و الگوی چند جمله‌ای کامل
۸-۴	۳-۴-۴- معایب روش الگوی کامل
۸-۴	۵-۴- الگوی استقلال مرتبه اول
۱۰-۴	۶-۴- الگوی بهادر
۱۲-۴	۱-۶-۴- برآورد پارامترها
۱۲-۴	۷-۴- روش I آمین نزدیکترین همسایه
۱۳-۴	۱-۷-۴- روش $I-NN$ برای متغیرهای پیوسته
۱۵-۴	۸-۴- روش الگوی لجستیک
۱۵-۴	۹-۴- استفاده از روشهای ممیزی گسسته برای متغیرهای پیوسته
۱۵-۴	۱۰-۴- سایر روشهای ممیزی گسسته
۱۷-۴	۱۱-۴- خروجیهای کامپیوتری
۴۴-۴	منابع و مراجع فصل چهارم

فصل پنجم: درخت کلاس بندی

۱-۵	۱-۵- مقدمه
۳-۵	۲-۵- کلاس بندهای با ساختار درختی
۵-۵	۳-۵- ساختن کلاس بند درختی
۶-۵	۱-۳-۵- انتخاب تقسیم‌ها
۶-۵	۱-۱-۳-۵- پرسشهای استاندارد
۷-۵	۲-۱-۳-۵- نیکویی تقسیم
۱۰-۵	۳-۱-۳-۵- ناخالصی درخت و ارتباط آن با نیکویی تقسیم
۱۱-۵	۲-۳-۵- قاعده توقف تقسیم

- ۱۱-۵ ۳.۳.۵- قاعده تخصیص شماره کلاس
- ۱۲-۵ ۱.۳.۳.۵- قاعده تخصیص با در نظر گرفتن زیانهای ناشی از رده بندی نادرست
- ۱۴-۵ ۴.۵- ترکیب متغیرها
- ۱۵-۵ ۵.۵- مقادیر گمشده
- ۱۵-۵ ۶.۵- پیرایش درخت کلاس بندی
- ۱۶-۵ ۷.۵- مزایا و معایب روش درختی و یک روش پیشنهادی
- ۱۸-۵ ۸.۵- درخت کلاس بندی با استفاده از تحلیل ممیزی تعمیم یافته
- ۱۸-۵ ۹.۵- الگوریتم سریع برای درختهای کلاس بندی
- ۱۹-۵ ۱.۹.۵- قاعده تقسیم به صورت ترکیب خطی و یک روش پیشنهادی دیگر
- ۲۰-۵ ۲.۹.۵- قاعده تقسیم یک متغیره
- ۲۱-۵ ۱۰.۵- متغیرهای ناترتیبی
- ۲۲-۵ ۱۱.۵- قاعده توقف و تخصیص شماره کلاس
- ۲۳-۵ ۱۲.۵- خروجیهای کامپیوتری
- ۳۸-۵ منابع فصل پنجم
- ۱.۶- مقاله تحلیل ممیزی آماری و روشهایی برای آن
- ۲.۶- نتایج حاصل و مقایسه آنها
- ۱.۷- برنامه های کامپیوتری (SPSS/PC+ , S-PLUS , SAS)

فصل اول

تابع ممیز خطی

۱.۱ - مقدمه

مسئله اصلی تحلیل ممیزی، تخصیص یک واحد آماری، به یکی از دو (یا چند) جامعه مجزا و براساس مقدار مشاهده شده x از آن است. در بعضی از موارد، اطلاعات نسبتاً کاملی از توزیع x در دو جامعه در دسترس است، و در بسیاری از موارد دیگر، اطلاعات مربوط به توزیع x از نمونه‌های بالنسبه کوچکی که از هر یک از جوامع π_1 یا π_2 (و در حالت کلی از $\pi_1, \pi_2, \dots, \pi_p$ یا π_G) اختیار می‌شوند، حاصل می‌گردد.

برخی از محققین، با تحلیل ممیزی به عنوان روشی برخورد می‌کنند که برای توصیف و آزمون تفاوت‌های بین جوامع به کار می‌رود. این آزمونها، با آزمونهای برابری میانگین‌های دو جامعه (و در حالت G جامعه، با آزمونهای برابری میانگین‌های چند جامعه) نرمال یکسان‌اند. به طور کلی می‌توان تحلیل ممیزی را به صورت زیر تعریف نمود:

۲.۱ - تعریف تحلیل ممیزی [1]

G جامعه $\pi_1, \pi_2, \dots, \pi_p$ و داده‌هایی از هر یک از آنها، به صورت زیر داده شده است:

$$\pi_1: x_1^{(1)}, x_2^{(1)}, \dots, x_{n_1}^{(1)} \quad ; \quad (n_1 \text{ حجم نمونه اصلی جامعه اول})$$

$$\pi_2: x_1^{(2)}, x_2^{(2)}, \dots, x_{n_2}^{(2)} \quad ; \quad (n_2 \text{ حجم نمونه اصلی جامعه دوم})$$

...

$$\pi_G: x_1^{(G)}, x_2^{(G)}, \dots, x_{n_G}^{(G)} \quad (n_G \text{ حجم نمونه اصلی جامعه } G \text{ ام})$$

که تعلق این نمونه‌ها به جوامع متناظر محرز بوده و ما آنها را «نمونه‌های اصلی^۱» می‌نامیم. همچنین داده‌های افرادی^۲ در دست است که متعلق بودن آنها به جوامع بالا از قبل مشخص نیست، و آنها را «نمونه‌های آزمون^۳» می‌گوییم. لازم به ذکر است که برداری $k \times 1$ است.

در تحلیل ممیزی معمولاً فرض می‌شود که نمونه‌های اصلی به درستی رده‌بندی می‌شوند، و متغیر یامتغیرهایی وجود دارند که با تحلیل روی آنها می‌توانیم هویت نمونه‌های آزمون را تعیین نموده و با کمترین میزان خطا، آنها را فقط به یکی از این G جامعه تخصیص دهیم. این نوع دسته‌بندی را که با توجه به گروه‌های شناخته شده انجام می‌گیرد، «تحلیل ممیزی^۳» یا «رده‌بندی با الگو^۴» می‌نامیم.

۱.۲.۱ - مراحل تحلیل ممیزی

به طور کلی مسائل رده‌بندی آماری در چهار مرحله، به شرح ذیل، انجام می‌گیرد:

۱- مشخص نمودن جوامع یا «تحلیل خوشه‌ای^۵»: در این قسمت با استفاده از داده‌ها، جوامع طوری مشخص می‌شوند که افراد یک گروه، از جهاتی شباهت و همانندی بیشتری نسبت به سایر دسته‌ها داشته باشند. لازم به ذکر است که در تحلیل ممیزی فرض می‌کنیم که G جامعه از قبل مشخص و تعریف شده‌اند.

۲- جمع‌آوری اطلاعات: صفات A_1, \dots, A_k را که می‌توانند کمی و یا کیفی باشند، به صورت بردار

$$\underline{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}$$

اندازه‌گیری می‌نمائیم. متغیرهای x_1, \dots, x_k را «متغیرهای پیشگو^۶» می‌نامند.

۳- اعمال الگوریتم‌های آماری: این مرحله شامل دو قسمت است، که عبارتند از:

الف. پیدا کردن قاعده‌ای که با استفاده از نمونه‌های اصلی، ساخته شده و تفاوت‌های G جامعه را به

1. Training samples یا Design set.
2. Test samples یا Test set.
3. Discriminant Analysis.
4. Classification with a teacher.
5. Clustering Analysis یا classification without a teacher.
6. Predictor variables.

بهترین نحو ممکن نشان می‌دهد.

ب. تخصیص نمونه‌های آزمون به یکی از جوامع که با استفاده از قاعده ساخته شده در مرحله الف انجام می‌گیرد. در این قسمت هدف اصلی رده‌بندی صحیح با بیشترین دقت است.

۴- ارائه و تفسیر صحیح نتایج: واضح است که دقت در انجام هر یک از مراحل فوق، می‌تواند صحت نتایج مراحل دیگر را به طور قابل ملاحظه‌ای تحت تأثیر قرار دهد. علاوه بر مراحل بالا می‌توان نکات دیگری را که در کاربردهای علمی تحلیل ممیزی مورد توجه قرار می‌گیرند، به صورت زیر فهرست نمود:

۱- قاعده تفکیک کننده، تا چه اندازه خوب عمل می‌کند؟

۲- قاعده تفکک کننده، تا چه اندازه توانایی تخطی از فرضهای معمول را دارد؟

۳- چه متغیرهایی را می‌بایست، جهت استفاده در قاعده تفکیک انتخاب نمود؟ [2]

۲.۲.۱- ارائه چند زمینه کاربردی

۱- در زمینه پزشکی: در تشخیص سرطان ریه، آزمایشاتی بر روی بافتهای ریه انجام می‌گیرد، و بر مبنای نتایج حاصل، بیمار به یکی از دو دسته سرطانی یا غیر سرطانی تخصیص می‌یابد. این متغیرها (مقادیر حاصل از آزمایشات) به علت‌هایی همچون مشکلات عملی بافت برداری از ریه‌ها و ناراحتیهایی که برای بیمار سبب می‌شوند، به سادگی نمی‌توانند مورد استفاده قرار گیرند. و در صورتیکه محقق صرفاً از روشهای بالینی برای تشخیص استفاده نماید، مجبور خواهد بود که تنها بیمارانی را که احتمال زیادی به سرطانی بودن آنها می‌دهد، مورد بررسی قرار دهد. بنابراین برای ممیزی، باید از متغیرهای دیگری که به طور ساده‌تری جمع‌آوری شده و در زمان بروز بیماری قابل استفاده باشند، سود جست. [2] در نتیجه اگر تشخیص یک بیماری به سرعت انجام گیرد، مداوای ناشی از این تشخیص می‌تواند تعداد بیماران را کاهش داده و در نتیجه می‌توان با به کارگیری همه جانبه روشهای ممیزی، به سالمتر شدن افراد جامعه کمک کرد.

۲- در زمینه فرهنگی: برای تعیین رشته یک دانش‌آموز در کنکور می‌توان از متغیرهای متعددی

مانند رتبه‌های هر آزمون، معدل دیپلم، سن، جنس، علائق شخصی، و ... به منظور تعیین مناسب‌ترین رشته