التحليل التمييزي(Discriminant Analysis)

المقدمة

التحليل التمييزي هو تقنية إحصائية متقدمة تهدف إلى تصنيف الحالات أو العناصر بناءً على مجموعة من المتغيرات المستقلة .يُعتبر هذا الأسلوب أداة قوية في العديد من المجالات مثل الطب، التسويق، المالية، والتعليم، حيث يتم استخدامه لتحديد الفئة التي تنتمي إليها حالة معينة بناءً على خصائصها الكمية.

في هذا الدرس، سنقدم إطارًا نظريًا متكاملًا للتحليل التمييزي مع التركيز على التفاصيل الرياضية والإحصائية المرتبطة به .سنتناول الأساسيات النظرية بشكل موسع، بما في ذلك النماذج الرياضية، الفرضيات الأساسية، وطرق بناء النموذج التمييزي.

1. الإطار النظري العام للتحليل التمييزي

أ .تعريف التحليل التمييزي

التحليل التمييزي هو أداة إحصائية تهدف إلى بناء نموذج رياضي يمكنه التمييز بين فئات مختلفة بناءً على مجموعة من المتغيرات المستقلة .يتم استخدام هذا النموذج لتصنيف الحالات الجديدة بناءً على قيم هذه المتغيرات.

ب .الغرض الأساسي

- تحديد العلاقة بين المتغيرات المستقلة والفئات المختلفة.
- بناء دالة تمييزية تساعد في تصنيف الحالات الجديدة بشكل دقيق.

ج .أنواع التحليل التمييزي

1. **التحليل التمييزي الخطي:(Linear Discriminant Analysis - LDA)**
   - يستخدم عندما تكون العلاقات بين المتغيرات والفئات خطية.
   - يعتمد على افتراض أن البيانات داخل كل فئة تتبع توزيعًا طبيعيًا متعدد المتغيرات وأن مصفوفة التباين والتباينات مشتركة بين الفئات.
2. **التحليل التمييزي غير الخطي:(Quadratic Discriminant Analysis - QDA)**
   - لا يفرض افتراضات صارمة حول مصفوفة التباين والتباينات المشتركة.
   - يسمح بوجود علاقات غير خطية بين المتغيرات والفئات.
3. **التعميمات الحديثة:**
   - التحليل التمييزي باستخدام طرق تقليل الأبعاد مثل Partial Least Squares Discriminant Analysis (PLS-DA).
   - الأساليب القائمة على تعلم الآلة مثل الشبكات العصبية والأشجار التقريرية.

2. النموذج الرياضي للتحليل التمييزي

أ .الفرضيات الأساسية

لتطبيق التحليل التمييزي، يجب أن تتحقق الافتراضات التالية:

4. **توزيع طبيعي:**
   - البيانات داخل كل فئة تتبع توزيعًا طبيعيًا متعدد المتغيرات.
   - إذا لم تكن البيانات تتبع توزيعًا طبيعيًا، قد يكون من الضروري استخدام تحويلات رياضية (مثل اللوغاريتم) لتحسين التوزيع.
5. **مصفوفة التباين والتباينات المتساوية (LDAفقط):**
   - التباينات (variances) ومصفوفة التباينات والانحرافات المشتركة (covariance matrix) متساوية بين الفئات.
   - إذا كانت التباينات غير متساوية، يتم استخدام QDA بدلاً من.LDA
6. **خطية العلاقة:**
   - العلاقة بين المتغيرات والفئات خطية.
   - إذا كانت العلاقة غير خطية، فإن QDA أو أساليب أخرى قد تكون أكثر ملاءمة.
7. **استقلال المتغيرات:**
   - المتغيرات المستقلة مستقلة تمامًا أو شبه مستقلة.
   - إذا كان هناك ترابط كبير بين المتغيرات، قد يؤدي ذلك إلى مشاكل في بناء النموذج.

ب .دالة التمييز الخطية

دالة التمييز الخطية هي نموذج رياضي يأخذ الشكل التالي:

$$D(X) = w_1X_1 + w_2X_2 + \cdots + w_pX_p + w_0$$

حيث:

- $X_1, X_2, \dots, X_p$: المتغيرات المستقلة.
- $w_1, w_2, \dots, w_p$: معاملات المتغيرات.
- w_0: الثابت.

ج .كيفية بناء دالة التمييز

.1**حساب المتوسط لكل فئة:**

يتم حساب المتوسط لكل متغير مستقل داخل كل فئة ($Y = 0$ و$Y = 1$):

$$\bar{X}_{i,j} = \frac{\sum_{k=1}^{n_j} X_{i,k}}{n_j}$$

حيث:

- i: يشير إلى المتغير (X_1 أو X_2).
- j: يشير إلى الفئة ($Y = 0$ أو $Y = 1$).
- n_j: عدد الحالات في الفئة j.

2. **حساب الفروق بين المتوسطات:**

يتم حساب الفرق بين المتوسطات لكل متغير بين الفئات:

$$w_i = \bar{X}_{i,1} - \bar{X}_{i,0}$$

3. **حساب الثابت:** w_0

يتم حساب الثابت باستخدام الصيغة التالية:

$$w_0 = -\frac{1}{2}\left(\sum_{i=1}^{p} \bar{X}_{i,1}^2 - \sum_{i=1}^{p} \bar{X}_{i,0}^2\right)$$

د .تصنيف الحالات الجديدة

لتصنيف حالة جديدة، يتم حساب قيمة الدالة التمييزية باستخدام قيم المتغيرات المستقلة الخاصة بها:

$$D(X) = w_1 X_1 + w_2 X_2 + \cdots + w_p X_p + w_0$$

ثم يتم اتخاذ القرار بناءً على عتبة التصنيف:

- إذا كانت: $D(X) > 0$ الحالة تنتمي إلى الفئة $Y = 1$.
- إذا كانت: $D(X) \leq 0$ الحالة تنتمي إلى الفئة $Y = 0$.

3. التفسير الإحصائي للتحليل التمييزي

أ .المسافة التمييزية

دالة التمييز يمكن اعتبارها مقياسًا للمسافة بين الحالة الجديدة وكل فئة .الفكرة الأساسية هي اختيار الفئة الأقرب إلى الحالة الجديدة بناءً على المسافة التمييزية.

ب .الاحتمالية الشرطية

يمكن تفسير التحليل التمييزي كتقدير للاحتمالات الشرطية:

$$P(Y = j \mid X) \propto f_j(X) \cdot P(Y = j)$$

حيث:

- $f_j(X)$:الكثافة الاحتمالية للحالة X ضمن الفئة j.
- $P(Y = j)$:الاحتمال الأولي للفئة j.

ج .العلاقة مع التحليل العاملي

التحليل التمييزي يرتبط ارتباطًا وثيقًا بالتحليل العاملي، حيث يمكن اعتباره طريقة لتقليل الأبعاد عن طريق إيجاد مكونات تمييزية رئيسية تفصل بين الفئات.

4. النماذج الرياضية المتقدمة

أ .التحليل التمييزي الخطي:(LDA)

فيLDA ، يتم افتراض أن مصفوفة التباين والتباينات مشتركة بين الفئات .يتم بناء النموذج باستخدام الوسائل التالية:

1. **إيجاد متجهات التمييز الرئيسية :**يتم حساب متجهات تمييزية باستخدام مصفوفة التباين والتباينات المشتركة بين الفئات.
2. **بناء دالة التمييز :**يتم استخدام المتوسطات والفروق بين الفئات لبناء دالة تمييزية خطية.

ب .التحليل التمييزي غير الخطي:(QDA)

فيQDA ، يتم السماح بمصفوفات تباين وتباينات مختلفة لكل فئة .النموذج يتطلب بيانات أكبر ويكون أكثر تعقيدًا من.LDA

5. حدود التحليل التمييزي

أ .الافتراضات القوية

- يتطلب LDA افتراضات صارمة مثل التوزيع الطبيعي ومصفوفة التباين والتباينات المشتركة.
- قد لا يكون فعالًا عند وجود علاقات غير خطية بين المتغيرات والفئات.

ب .حساسية النموذج للبيانات

- النموذج يعتمد بشكل كبير على جودة البيانات وحجم العينة.
- قد يؤدي وجود قيم شاذة (outliers) إلى تأثير كبير على دقة النموذج.

ج .التوازن بين الفئات

- إذا كانت الفئات غير متوازنة (أي أن أحدها يحتوي على عدد أكبر بكثير من الحالات)، فقد يؤثر ذلك على دقة التصنيف.

6. التطبيقات النظرية

أ .في مجال التسويق

- تصنيف العملاء بناءً على سلوكهم الشرائي (مستهدفون وغير مستهدفين).

- تحديد المنتجات التي قد يشتريها العملاء بناءً على تاريخهم السابق.

ب .في مجال الطب

- تشخيص الأمراض بناءً على المؤشرات البيولوجية مثل مستوى السكر في الدم أو ضغط الدم.
- تصنيف المرضى بناءً على مخاطر الإصابة بأمراض معينة.

ج .في مجال المالية

- تقييم المخاطر الائتمانية للعملاء بناءً على مؤشرات مثل الدخل الشهري والديون السابقة.
- تصنيف الشركات بناءً على أدائها المالي.

د .في مجال التعليم

- تحديد الطلاب الذين قد يحتاجون إلى دعم إضافي بناءً على أدائهم الأكاديمي.
- تصنيف الطلاب بناءً على احتمالية نجاحهم في اختبارات معينة.