المحور الأول: الانحدار اللوجستي

1. مقدمة إلى الانحدار اللوجستي

الانحدار اللوجستي هو حجر الزاوية في النمذجة الإحصائية، خاصة في السيناريوهات التي يكون فيها الناتج ثنائيًا)مثل: نجاح/فشل، نعم/لا.(يربط هذا الأسلوب بين الانحدار الخطي ومهمات التصنيف من خلال التنبؤ بالاحتمالات بدلاً من القيم المستمرة. تضمن الدالة اللوجستية أن تكون التنبؤات محصورة بين 0 و1، مما يجعلها مناسبة للتفسير الاحتمالي.

التطبيقات الرئيسية:

- **الحفاظ على العملاء**: التنبؤ بما إذا كان العميل سيترك الشركة أم سيظل وفيا لها.
- **الرعاية الصحية**: تشخيص الأمراض بناءً على بيانات المرضى.
- **التسويق**: تحديد المشترين المحتملين من البيانات الديموغرافية.
- **المالية**: تقييم مخاطر الائتمان(تخلف/عدم تخلف).

2. الدالة اللوجستية

الدالة السينية، المعرّفة كالتالي:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

هي جوهر الانحدار اللوجستي، تقوم هذه الدالة بتحويل أي قيمة حقيقية إلى قيمة بين 0 و1، تمثل احتمال الفئة الإيجابية $Y = 1$)).

لماذا يتم استخدام الدالة السينية؟

- **التفسير الاحتمالي**: يمكن تفسير المخرجات مباشرة كاحتمالات.
- **انتقال سلس**: المنحنى ينتقل بسلاسة بين 0 و1، مما يتجنب التغيرات المفاجئة.
- **البساطة الرياضية**: مشتقاتها بسيطة، مما يساعد في خوارزميات التحسين.

3. افتراضات الانحدار اللوجستي

فهم الافتراضاتالتي يقوم عليها الانحدار اللوجستي أمر بالغ الأهمية لتطبيقه بشكل صحيح:

1. **الناتج الثنائي**: يجب أن تكون المتغير التابع ثنائيًا (مثل: 0/1)

2. **خطية لوغاريتم الاحتمالات**: العلاقة بين المتغيرات التنبؤية ولوغاريتم الاحتمالات للنتيجة يجب أن تكون خطية.
3. **الاستقلالية**: يجب ألا تؤثر الملاحظات على بعضها البعض(لا يوجد ارتباط ذاتي).
4. **عدم وجود تعدد الخطية**: يجب ألا تظهر المتغيرات التنبؤية ارتباطات عالية، مما قد يؤدي إلى زعزعة استقرار تقديرات المعاملات.
5. **حجم عينة كبير**: يعمل الانحدار اللوجستي بشكل أفضل مع مجموعات بيانات أكبر، مما يضمن تقديرات موثوقة للمعلمات.

4. صياغة النموذج: من الخطي إلى اللوجستي

يبدأ نموذج الانحدار اللوجستي بمفهوم **لوغاريتم الاحتمالات** :(log)

$$\log(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

يعبر هذا المعادلة عن اللوغاريتم الطبيعي لاحتمالات $Y = 1$ كمجموعة خطية من المتغيرات التنبؤية. بحل المعادلة بالنسبة لـ P، نحصل على:

$$P = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_n X_n}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_n X_n}}$$

تؤكد هذه الصياغة أن P يبقى ضمن المجال $[0, 1]$.

5. تقدير المعلمات: الاحتمال الأقصى

يُستخدم تقدير الاحتمال الأقصى (MLE) لتقدير المعاملات (β) التي تعظم احتمال ملاحظة البيانات المعطاة. دالة الاحتمال هي:

$$L(\beta) = \prod_{i=1}^{N} P_i^{Y_i} (1 - P_i)^{1-Y_i}$$

أخذ اللوغاريتم الطبيعي يبسط العملية من ضرب إلى جمع:

$$\ln L(\beta) = \sum_{i=1}^{N} [Y_i \ln P_i + (1 - Y_i) \ln(1 - P_i)]$$

تُستخدم تقنيات التحسين مثل **نيوتن-رافسون**، أو الانحدار التدريجي، أو طرق شبه نيوتن لإيجاد القيم المثلى لـ β.

6. تقييم النموذج

تقييم نموذج الانحدار اللوجستي يتضمن تقييم أدائه التنبؤي وقدرته على التفسير.

المقاييس الشائعة:

1. **مصفوفة الارتباك**: جدول يلخص الإيجابيات الحقيقية (TP)، السلبيات الحقيقية (TN)، الإيجابيات الكاذبة (FP)، والسلبيات الكاذبة .(FN)
2. **الدقة**: نسبة التنبؤات الصحيحة:

$$\text{الدقة} = \frac{TP + TN}{TP + TN + FP + FN}$$

3. **الدقة التنبؤية**: نسبة التنبؤات الإيجابية التي كانت صحيحة:

$$\text{الدقة التنبؤية} = \frac{TP}{TP + FP}$$

4. **الاستدعاء(الحساسية)**: نسبة الإيجابيات الحقيقية التي تم تحديدها بشكل صحيح:

$$\text{الاستدعاء} = \frac{TP}{TP + FN}$$

5. **ROC-AUC**: يقيس التوازن بين معدل الإيجابيات الحقيقية (الحساسية) ومعدل الإيجابيات الكاذبة (الخصوصية) عبر عتبات مختلفة.

7. تفسير المعاملات

يمثل كل معامل (β) في الانحدار اللوجستي التغيير في لوغاريتم الاحتمالات لكل زيادة بوحدة واحدة في المتغير التنبؤي المقابل. يؤدي رفع المعامل إلى الأس إلى الحصول على **نسبة الاحتمالات**:

$$\text{نسبة الاحتمالات} = e^{\beta}$$

- **نسبة الاحتمالات :1** < المتغير التنبؤي يزيد من احتمال $Y = 1$.
- **نسبة الاحتمالات :1** > المتغير التنبؤي يقلل من احتمال $Y = 1$.