استخدام الأساليب اللامعلمية لتقدير الدالة التمييزية

م. طاهر ريسان دخيل قسم الإحصاء/ جامعة القادسية/كلية الإدارة والاقتصاد
م .غسق فاضل عبد الرزاق /الجامعة المستنصرية / قسم الإحصاء / كلية الإدارة والاقتصاد
م. زينب يوسف داود/ الجامعة المستنصرية /كلية الإدارة والاقتصاد / قسم الإحصاء

الخلاصة

هناك بعض الافتراضات التي يجب أن تتوفر في المتـــــغيرات العشوائية بهدف استخدام الطرائـــــق المعلمية في تقدير الدالة الـــــــتمييزية ومن هذه الافتـــــراضات أن يكون توزيع هذه المتغيرات هو توزيعاً طبيعياً ولكن هذا الافتــراض قد لا يتحقق مما يتطلب البحث عن طرائق بديلة لتـــــــــقدير الدالة ومن هذه الطرائق أسلوب التقدير اللامعلمي المسمى بتقدير Kernelأو طريقة Kernelوالتي تعتمد على البيانات في الحصول على تقدير للدالة مما يجنبنا المرور بتقدير المعلمات للدالة التميزية ،ففي هذا البحث تم استخدام أسلوب Kernel ولثلاث دوال هي دالة Epanchinicove وGaussion وSin ومقارنة النتائج مع الطريقة المعلمية التقليدية وذلك بالاعتماد على مصادر خطأ التصنيف في الدالة التمييزية .

Abstract

Some assumptions must available in random variables to allow us to use the parametric methods in discriminate function estimation. One of these assumptions is that the distribution of the variables are normal distribution ,but sometimes this is not achievement so we must seek for an alternative methods to estimate the discriminant function ,one of these methods is the nonparametric estimation procedure which called Kernel estimation or Kernel method which depends on data to get the estimation. In this paper we use Kernel procedure with three functions, Epanchinikov, Gaussian & Sine functions and compare results with the parametric method by using misclassification as criterion of comparing.

المقدمة [5][1]

أن الهدف من الدالة التمييزية هو تصنيف المشاهدات إلى مجموعتين أو أكثر بحيث يكون التباين بين تلك المجاميع أعظم ما يمكن بينما يكون التباين داخل كل مجموعة اقل ما يمكن وذلك من خلال بناء معادلة ذات تركيبة خطية وبالتالي يمكن استخراج الطرائق المعلمية في تقدير تلك الدالة وهذا يمكن أن يتم في حالة وجود الفروض الأساسية لهذا التحليل ولكن عندما لا يتوفر ذلك يمكن الاتجاه إلى الأساليب اللامعلمية،حيث اتجهت البحوث حديثا إلى دراسة طرائق التقدير اللامعلمي والتقدير اللامعلمي لا يعني أن النموذج لا يحوي على معلمات وإنما هو طريقة لا تعتمد على المعلمات في التقدير وإنما تعتمد على البيانات مباشرة ،ومن تلك الأساليب أسلوب Kernel في التقدير والذي يستخدم آلية خاصة في عملية التقدير .

لقد اقترحت طريقة Kernel في التقدير ابتدءاً من قبل الباحث Osenblatt [5] عام ١٩٥٦ وطورت هذه الطريقة من قبل الباحث Parzen عام ١٩٦٢ وهناك بعض البحوث التي قدمت ضمن التقدير اللامعلمي والتي تضمنت تقديرات Kernel ومنه البحث الذي قدم من قبل الباحث Silverman عام ١٩٨٦ والذي تضمن استخدام دوال Kernelفي تقدير الدوال التميزية .

هدف البحث

يهدف هذا البحث إلى المقارنة بين الطريقة المعلمية وطريقة Kernel اللامعلمية في استخدام عدة دوال وذلك من خلال استخدام المحاكاة في حالة عدم تحقق الفروض الأساسية للدالة التمييزية وقد استخدم معيار خطأ التصنيف كأساس للمقارنة .

الدالة الخطية للتصنيف [4][1]

يعتبر أسلوب التصنيف احد الأساليب المهمة ضمن متعدد المتغيرات وذلك بهدف معرفة إلى أي مجتمع من المجتمعات تعود مشاهدة ما والمبدأ الأساس الذي يستند علية هذا التصنيف هو بتقليص خطا التصنيف أو سوء التصنيف وهو أن نضيف مشاهدة معينة بأنها تعود للمجموعة الأولى مثلاً بينما هي في حقيقة الأمر تعود إلى المجموعة الثانية وهذا المعيار هو معيار مهم فمن خلاله يمكن معرفة قوة الدالة التمييزية أو التصنيفية فكلما كان عدد المشاهدات المصنفة خطا قليل فان تلك الدالة جيدة وقادرة على فرز المشاهدات إلى مجتمعاتها الأصلية بصورة صحيحة والعكس بالعكس.

ولغرض شرح الأساس النظري لأسلوب التمييز لنفرض أن هناك عينتين عشوائيتين هما n1 وn2 للمجموعة الأولى والثانية على التوالي والتي تتكون من p من المتغيرات العشوائية ويفترض أن هاتين العينتان سحبتا من مجتمعين يتوزعان توزيعاً طبيعياً

بمتوسط μ_1 و μ_2 للمجموعة الأولى والثانية على التوالي وبتباين Σ للمجموعتين ومن هاتين العينتين يتم حساب $\underline{\overline{Y}_1}$ و $\underline{\overline{Y}_2}$ واللذان يمثلان متجها المتوسطات المقدرة من العينة للمجموعة الأولى والثانية على التوالي وكذلك حساب S_{pl} والتي تمثل مصفوفة التباين المشترك المقدر ،وبعدها يمكن حساب الدالة التمييزية المقدرة وبالشكل الأتي

$$Z = a'\underline{Y} = (\underline{\overline{Y}_1} - \underline{\overline{Y}_2})'(S_{pl})^{-1}Y \ldots\ldots\ldots 1$$

حيث إن $\underline{Y}$ هو متجه المشاهدات الجديدة والتي نرغب في تصنيفها إما إلى المجموعة الأولى أو الثانية وحسب المعلومات التي تحملها.

ولغرض تحديد إن $\underline{Y}$ تعود لإحدى المجموعتين فأننا نعتمد بذلك على أن Z في المعادلة ١ هل هي قريبة إلى المتوسط $\overline{Z}_1$ أو $\overline{Z}_2$ حيث إننا نستخدم بيانات المجموعة الأولى $\overline{Y}_{il}$ لتحديد القيم $Z_{11}, Z_{12}, Z_{13}, \ldots Z_{1n_1}$ ومن ثم حساب $\overline{Z}_1$ والذي يمكن أيضا أن يحسب وفق الصيغة الآتية

$$\overline{Z}_1 = a'\underline{\overline{Y}}_1 = (\underline{\overline{Y}}_1 - \underline{\overline{Y}}_2)'(S_{pl})^{-1}\underline{\overline{Y}}_1 \ldots\ldots\ldots 2$$

وبشكل مشابه يمكن حساب قيمة $\overline{Z}_2$ والتي تساوي $\overline{Z}_2 = a'\underline{\overline{Y}}_2$ وبالتالي فان المشاهدة تعود للمجموعة الأولى إذا كانت $\overline{Z} = a'Y$ قريبة الى المجموعة $\overline{Z}_1$ اكثر من $\overline{Z}_2$ وتعود الى المجموعة الثانية اذا كانت قريبة من $\overline{Z}_2$ اكثر من $\overline{Z}_1$ او بشكل اخر فان المشاهدة تعود الى المجموعة الاولى اذا كانت $Z > \frac{1}{2}(\overline{Z}_1 + \overline{Z}_2)$ وحيث ان المسافة بين $\overline{Z}_1$ و $\overline{Z}_2$ هي نفس المسافة بين $\underline{\overline{Y}}_1$ و $\underline{\overline{Y}}_2$ اذ يمكن كتابة قاعدة التصنيف وذلك بالاعتماد على قيم Y بالشكل الاتي

$$\frac{1}{2}(\overline{Z}_1 + \overline{Z}_2) = \frac{1}{2}(\underline{\overline{Y}}_1 - \underline{\overline{Y}}_2)'(S_{pl})^{-1}(\underline{\overline{Y}}_1 + \underline{\overline{Y}}_2) \ldots\ldots\ldots 3$$

وبالتالي فان قاعدة التصنيف تصبح بالشكل الاتي

$$a'Ya = (\underline{\overline{Y}}_1 + \underline{\overline{Y}}_2)'S_{pl}^{-1}Y > \frac{1}{2}(\underline{\overline{Y}}_1 - \underline{\overline{Y}}_2)'S_{pl}^{-1}(\underline{\overline{Y}}_1 + \underline{\overline{Y}}_2) \ldots\ldots\ldots 4$$

وفي هذه الحالة فان المشاهدة تصنف الى المجموعة الاولى بينما تصنف المشاهدة الى المجموعة الثانية اذا كانت

$$a'Ya = (\underline{\overline{Y}}_1 - \underline{\overline{Y}}_2)'S_{pl}^{-1}Y < \frac{1}{2}(\underline{\overline{Y}}_1 - \underline{\overline{Y}}_2)'S_{pl}^{-1}(\underline{\overline{Y}}_1 + \underline{\overline{Y}}_2) \ldots\ldots\ldots 5$$

خطا التصنيف Miss Classification [5][1]

آن خطا التصنيف يكمن في ان الدالة تصنف مشاهدة معينة في المجموعة الأولى وهي في حقيقة الآمر تعود الى المجموعة الثانية او العكس أي أنها تصنف مشاهدة معينة الى المجموعة الأولى وهي تعود الى المجموعة الثانية ويمكن توضيح ذلك بالجدول الاتي

المجموعة الحقيقية	عدد المشاهدات	المجموعة المتنبأ بها	
		١	٢
١	n_1	n_{11}	n_{12}
2	n_2	n_{21}	n_{22}

من خلال الجدول أعلاه يمكن إيجاد نسبة خطا التصنيف للدالة التميزية المقدرة ولنرمز لها بالرمز E

$$\left.\begin{aligned} E &= \frac{n_{12}+n_{21}}{n_1+n_2} \\ E &= \frac{n_{12}+n_{21}}{n_{11}+n_{12}+n_{21}+n_{22}} \end{aligned}\right] \quad \ldots\ldots\ldots 6$$

وكذلك يمكن حساب نسبة التصنيف الصحيح بالشكل الاتي

$$C = \frac{n_{11}+n_{22}}{n_1+n_2} \quad \ldots\ldots\ldots 7$$

حيث ان C يمثل نسبة التصنيف الصحيح

التصنيف بالاعتماد على دوال Kernel [6][5][4]

إن التصنيف يعتمد على الفروض الأساسية لقواعد التصنيف الخطية والتربيعية ومنها أن المتغيرات تتبع التوزيع الطبيعي وفي حالة عدم تحقق هذا الشرط فانه يمكن الانتقال إلى الطرائق اللامعلمية وأحدى هذه الطرائق تسمى طريقة Kernel وهنا يمكن إعطاء شرحاً عن طريقة Kernel ،فعلى افتراض وجود متغير Y والذي يمثل الدالة f(y) والتي يطلب تقديرها باستخدام العينة العشوائية $y_1, y_2, y_3, \ldots, y_n$ إن التقدير ل f(y) لأي نقطة (y_o) يمكن أن تعتمد على النسبة للنقاط ضمن الفترة $(y_o - h, y_o + h)$ فإذا رمزنا لعدد

النقاط ضمن الفترة المذكورة بالرمز $N(y_o)$ فستكون النسبة $N(\frac{y_o}{h})$ تقديرا إلى $p(y_0-h<y_0<y_0+h)$ والتي تكون مساوية بشكل تقريبي للمقدار $2h(y_o)$ ولذلك يمكن تقدير $f(y_o)$ باستخدام المعادلة الاتية

$$f(y_o)=\frac{N(y_o)}{2hn} \quad \text{.....8}$$

ويمكن وضع المقدار $\hat{f}(y_o)$ كدالة لكل قيم y_i في العينة وذلك من خلال الدالة الاتية

$$K(u)=\begin{cases}\frac{1}{2} & for\ |u\le 1|\\ 0 & for\ |u\le 1|\end{cases} \quad \text{.....9}$$

حيث أن الدالة أعلاه تسمى بالدالة المنتظمة Uniform function وبعد التعويض عن قيمة u بالمقدار $\frac{y_o-y_i}{h}$ لنحصل على معادلة

$$Ny_o=2\sum_{i=1}^{n}k(\frac{y_o-y_i}{h}) \quad \text{.....10}$$

لتصبح المعادلة رقم ٨ بالشكل الاتي

$$f(y_o)=\frac{1}{hn}\sum_{i=1}^{n}k(\frac{y_o-y_i}{h}) \quad \text{.....11}$$

وان المقدار $k(\frac{y_o-y_i}{h})$ يعني أنها تساوي نصف لآي نقطة y_i ضمن الفترة (y_o-h,y_o+h) وتساوي صفر لآي نقطة خارج هذه الفترة .

ويمكن كذلك استخدام دوال تابعة لأسلوب Kernel غير الدالة المنتظمة uniform function ومن تلك الدوال

$$k(u)=\frac{1}{\pi}\frac{\sin^2(u)}{u^2} \quad \text{.....12}$$

حيث تسمى هذه الدالة بالدالة الجيبية sine function .

والدالة التالية تسمى بالدالة الكاوسية (Gaussian function)

$$k(u)=\frac{1}{\sqrt{2\pi}}e^{-u^2/2} \quad \text{.....13}$$

اما الدالة التالية فتسمى بدالة Epanechnikov

$$k(u)=\begin{cases}\frac{3}{4}(1-u^2) & if \quad (|u|\le 1)\\ 0 & otherwise\end{cases} \quad \text{.....14}$$

حيث ان دوال kernel هذه لا تتضمن أي افتراضات حول دالة الكثافة f(y) أن عملية اختيار قيمة h التي تستخدم مع دوال kernel يجب أن تكون بحذر شديد ،إذ أن إذا كانت قيمة h صغيرة فان دالة $\hat{f}(y_o)$ ستكون لها قيمة عالية عند كل y_i بينما إذا كانت h كبيرة فان $\hat{f}(y_o)$ تكون غالباً منتظمة لذلك فان اختيار قيمةh يجب ان تعتمد على حجم العينة وذلك لتجنب عملية التمهيد الكبيرة او الصغيرة بحيث لو كان حجم العينة كبيراً فيجب اختيار قيمة h صغيرة والعكس بالعكس او يمكن اختيارعدة قيم لـــ h والمقارنة بينها من خلال قيمة الخطا .

ولغرض استخدام طريقة Kernel في تقدير الدالة التمييزية فانه يمكن تطبيقها على كل مجموعة وذلك لغرض تحديد $\hat{f}(y_o|g_1), \hat{f}(y_o|g_2)\cdots \hat{f}(y_o|g_k)$ اذ ان g_i ,i=1, 2, 3.....k تمثل المجاميع الغير معلومة ، لتصبح قاعدة التصنيف هي تعظيم للاحتمال الاتي

$$p_i(\hat{f}(y_o|g_i))$$

إذ أن g_i تمثل المجموعة i

بناء تجربة المحاكاة

لقد تم بناء تجربة المحاكاة باستخدام لغة البيسك بحيث يتم خرق الافتراض القائل بان المتغيرات الداخلة في عملية تقدير الدالة التميزية ذات توزيع طبيعي وذلك من خلال توليد مشاهدات تتبع التوزيع الاسي لأحد المتغيرات ومن ثم تقدير الدالة التميزية بالطريقة المعلمية التقليدية أولا وبعدها استخدام طريقة Kernel اللامعلمية وبثلاث دوال وهي دالة

Epanchinicove ودالة Gaussian والدالة الجيبية Sine function وذلك عند أحجام عينات مختلفة هي 250,200,150,100,50,30,10
وتكرار التجربة بواقع ١٠٠٠ مرة وملاحظة سلوك خطأ التصنيف لتتم المقارنة على أساسه. ويمكن تلخيص خطوات المحاكاة بالشكل الاتي
تحديد حجم العينة وكما مذكور سابقاً.
توليد مشاهدات لثلاث متغيرات ولمجموعتين بحيث يكون توزيع المتغير الثاني ذو توزيع أسي بالمعلمة $\lambda = \mathbf{0.3}$.
حساب تقديرا لدالة التميزية على وفق الطريقة المعلمية التقليدية وطريقة Kernel اللامعلمية وحسب الدوال المستخدمة والتي تم شرحها سابقاً.
تكرار التجربة ١٠٠٠ مرة .
حساب خطأ التصنيف وذلك بتحديد المشاهدات التي صنفت إلى المجموعة الخاطئة وحساب شبه الخطأ.

تحليل النتائج

١- عند استخدام دالة Epanchinicove نلاحظ بصورة عامة وعند استخدام هذه الدالة أن طريقة Kernel اللامعلمية تكون أفضل من الطريقة المعلمية التقليدية لتقدير دالة التصنيف فمثلاً عند حجم العينة n=10 نلاحظ أن الطريقة المعلمية التقليدية صنفت بعض المشاهدات خطأ وبنسبة 45% بينما صنفت طريقة Kernel اللامعلمية 40% من المشاهدات خطأ .
ونلاحظ أيضا أن خطأ التصنيف يتناقص عند زيادة حجم العينة ولكلتا الطريقتين مع بقاء الأفضلية لطريقة Kernel اللامعلمية ،هذا ما يوضحه الجدول رقم ١ في الملاحق فعند حجم العينة n=30 كانت نسبة خطأ التصنيف 16% , 28% بالنسبة للطريقة المعلمية التقليدية وطريقة Kernel اللامعلمية على التوالي وهكذا لجميع أحجام العينات المستخدمة حيث نلاحظ التقارب بين قيم خطأ التصنيف للطريقة المعلمية 7% بينما الطريقة اللامعلمية 6% .
٢- عند استخدام الدالة الكاوسية Gaussian function عند استخدام هذه الدالة فإننا نلاحظ أن نفس السلوك الذي تسلكه الدالة السابقة يمكن أن يسري على هذه الدالة إن 50% من المشاهدات صنفت خطأ عندما يتم استخدام الطريقة المعلمية التقليدية بينما 16% من المشاهدات صنفت خطأ استخدام طريقةKernel اللامعلمية وذلك حجم العينة

n=10 وهكذا بالنسبة لبقية أحجام العينات فعند حجم العينة n=250 نلاحظ ان خطأ التصنيف بلغت نسبته 6% لكلتا الطريقتين وبقية نسب خطأ التصنيف موضحة في جدول رقم ٢ وعموماً فان هناك تقارب بين قيم نسبة خطأ التصنيف عند استخدام هذه الدالة والدالة السابقة.

٣- عند استخدام الدالة الجيبية Sine function توضح قيم خطأ التصنيف في الجدول رقم ٣ أن 20% من المشاهدات صنفت خطأ عند استخدام الطريقة المعلمية بينما 32% من المشاهدات صنفت خطأ عند استخدام طريقة Kernel اللامعلمية وذلك عند حجم العينة n=10 وهكذا بالنسبة لبقية أحجام العينات المستخدمة ، حيث نلاحظ ان هذه الدالة مع الدالة السابقة (الدالة الكاوسية Gaussian function) افضل من دالة Epanchinicove وذلك بموجب النتائج التي أعطتها.

الاستنتاجات

١- كانت طريقة Kernel اللامعلمية افضل من الطريقة المعلمية التقليدية .

٢- ان دالتي Sine function و Gaussian function تعطي نتائج افضل من دالة Epanchinicove عند استخدام طريقة Kernel

٣- نلاحظ ان عند زيادة حجم العينة تقل نسبة خطأ التصنيف لكلتا الطريقتين .

المصادر

الجبوري ، شلال وعبد ، صلاح حمزة (٢٠٠٠) "تحليل متعدد المتغيرات "، الجامعة المستنصرية

Gavin, C (2005) “Efficient Cross -Validation of Kernel Fisher discriminant classifiers "Elsevier Scic,uk.

Hyunsoo Kim and Hyson park (2005) ,,Relations lips between support vector classifiers and Generalized linen discriminant analysis on support vectors " Technical Report university of Minnesota

Rencher, A (2002)" Methods of Multivariate Analysis " John Wiley & Sons ,Inc.

Shan ,L.and Tad E.C (2005) ' Nonlinear Kernel MSE. Methods for Cancer classification " Springer verlag Berlin Heidelberg p .p 975 – 984

Tao,X. , Y. ,Jieping , Qi, L. , C. , Vladimir &J. ,Ravi (2005)" Efficient Kernel Discriminate Analysis via QR decomposition " The Army high per furnace computing research center.

الملاحق

جدول رقم (١)

ويمثل قيم خطأ التصنيف عندما يتم استخدام دالة Epanchinicove

حجم العينة	الطريقة التقليدية	طريقة كيرنل
١٠	٠,٤٥	٠,٤٠
٣٠	٠,١٦	٠,٢٨
٥٠	٠,٢١	٠,١٦
١٠٠	٠,٢٣	٠,١٧
١٥٠	٠,١٠	٠,٠٩
٢٠٠	٠,٠٨	٠,٠٨
٢٥٠	٠,٠٧	٠,٠٦

جدول رقم (٢)

ويمثل قيم خطأ التصنيف عندما يتم استخدام دالة Gaussian

حجم العينة	الطريقة التقليدية	طريقة كيرنل
١٠	٠,٥٠	٠,١٦
٣٠	٠,٢١	٠,١٨
٥٠	٠,١٨	٠,١٥
١٠٠	٠,١٤	٠,١٤
١٥٠	٠,٠٩	٠,٠٦
٢٠٠	٠,٠٨	٠,٠٦
٢٥٠	٠,٠٦	٠,٠٦

جدول رقم (٣)

ويمثل قيم خطأ التصنيف عندما يتم استخدام الدالة الجيبية Sine

حجم العينة	الطريقة التقليدية	طريقة كيرنل
١٠	٠,٢٠	٠,٣٢
٣٠	٠,١٦	٠,١٣

٥٠	٠,١٢	٠,٢١
١٠٠	٠,٠٩	٠,٠٨
١٥٠	٠,٠٤	٠,٠٣
٢٠٠	٠,٠٣	٠,٠٥
٢٥٠	٠,٠٣	٠,٠٣