استخدام استدلال بيز في التصنيف مع تطبيق على أسلوب التحليل التمييزي

ذنون يونس ذنون الشكرجي
قسم التقنيات المالية والمصرفية
المعهد التقني / نينوى

الاستلام	القبول
2008 / ١١ / ٢٣	٢٠٠٩ / ٠٢ / ١٦

Summary

Bayesian Inference depends in general on the use of prior information about the unknown parameters $(\theta=\theta_1,\theta_2,...,\theta_n)$ to be estimated as such parameter random variables rather than constant quantities and add that prior information the sample information which is the information in the form of a prior information function and could we know that it is symptomatic of all the information And experiences on parameters and features to be reached in advance through analysis or control of those features. The distribution function of observations of the current sample (Y) depends on the parameters (θ) and denoted by $f(Y/\theta)$, in this research Bayesian Inference in the classification has been used with application on discriminant analysis approach for sample of students Technical Institute (Nineveh) the accounting department where the study included (85) students Two of the morning and evening study also included eight variables representative of school materials and taught in the accounting department (second stage), the results of the analysis showed that there are (17) observation was wrongly classified among (85) observation, and the proportion of misclassification (20%).

الملخص

يعتمد استدلال بيز بشكل عام على استخدام معلومات مسبقة حول المعـالم المجهولـة المطلوب تقديرها على اعتبار أن هذه المعالم متغيرات عشوائية وليست $(\theta=\theta_1,\theta_2,...,\theta_n)$

كميات ثابتة ويضاف إلى تلك المعلومات المسبقة معلومات العينة حيث تمثل تلك المعلومات على شكل دالة احتمالية أولية ويمكن أن نعرفها بأنها الدالة التي تمثل كل المعلومات والخبرات حول المعالم المراد تقديرها والتي تم التوصل اليها مسبقاً من خلال التحليل أو المراقبة لتلك المعالم. أما دالة التوزيع الاحتمالي لمشاهدات العينة الحالية (Y) فتعتمد على المعالم (θ) ويرمز لها $f(Y/\theta)$ ، تم في هذا البحث استخدام استدلال بيز في التصنيف مع تطبيق على أسلوب التحليل التمييزي لعينة من طلبة المعهد التقني (نينوى) قسم المحاسبة حيث شملت الدراسة (٨٥) طالباً وطالبة للدراستين الصباحية والمسائية كما اشتملت الدراسة على ثمانية متغيرات ممثلة بالمواد الدراسية والتي تدرس في القسم (المرحلة الثانية)، أظهرت نتائج التحليل هنالك (١٧) مشاهدة تم تصنيفها بصورة خاطئة من بين (٨٥) مشاهدة وان نسبة خطأ التصنيف بلغت (٢٠%).

المقدمة :-

أدت الاستخدامات العديدة والمتنوعة لعلم الإحصاء في مجالات الحياة المختلفة إلى اهتمام عدد من الباحثين بدراسة العلوم الإحصائية وبمرور الزمن نشأت عن هذه الدراسات المدرسة البيزية وتعتمد في أسلوبها وتحليلها واستنتاجاتها على المعلومات التي توفرها المشاهدات (العينات) فضلاً عن المعلومات التي تأتي من الاعتقاد الشخصي وتدعى بالمعلومات الأولية (Prior Information) وتتميز مدرسة بيز عن المدرسة التقليدية بأنها تعامل المعلمات في التوزيعات الاحتمالية بوصفها متغيرات عشوائية ولها توزيع احتمالي وأطلق على هذه المدرسة بمدرسة بيز نسبة إلى مؤسسها الإحصائي البريطاني توماس بيز (Thomas Bayes) (الرجبو، ٢٠٠٥).

هدف البحث

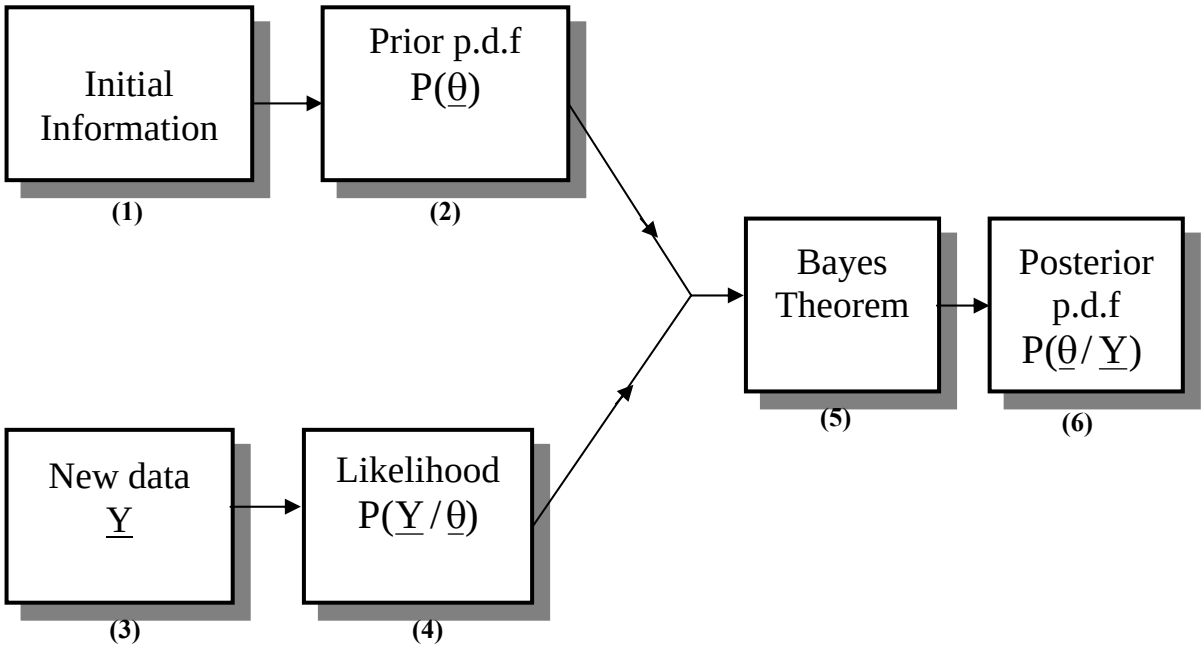
يهدف البحث إلى تطبيق استدلال بيز في التحليل التمييزي لتصنيف مجموعة من طلبة المعهد التقني / نينوى إلى مجموعتين (الصباحية والمسائية) من خلال إيجاد نتائج خطأ التصنيف لكلتا المجموعتين فضلاً عن إيجاد التوزيع الأولي بالاعتماد على حجم العينة المأخوذة من كل مجموعة وكذلك إيجاد التوزيع السابق واللاحق لكلتا المجموعتين.

نظرية بيز Bayes Theorem

تعتمد نظرية بيز على إستخدام معلومات مسبقة عن المعالم غير المعروفة وتعد هذه المعالم متغيرات عشوائية (Random Variables)، وبافتراض إن لهذه المعالم معلومات

مسبقة (Prior Information) التي يمكن صياغتها على شكل توزيع احتمالي يعرف بدالة الكثافة الاحتمالية السابقة (Prior P.d.f) ويجري التعرف على هذه المعلومات من بيانات وتجارب سابقة أو من النظرية التي تحكم تلك الظاهرة، وكذلك تعتمد نظرية بيز على معلومات العينة الحالية المتمثلة بدالة الأمكان (Likelihood Function) الخاصة بالمشاهدات، وعليه بدمج دالة الكثافة الاحتمالية للمعالم $P(\underline{\theta})$ مع دالة الأمكان $P(\underline{Y}/\underline{\theta})$، يتم الحصول على دالة الكثافة الاحتمالية اللاحقة للمعالم (Posterior p.d.f) وهي $P(\underline{\theta}/\underline{Y})$ ويمكن تلخيص ما تقدم بشكل مخطط وعلى وفق ما يأتي: (الرجبو، ٢٠٠٥)

شكل (١): مفهوم نظرية بيز



وباستخدام العمليات الرياضية الاعتيادية التي توضح الشكل السابق.

$$P(\underline{\theta}/\underline{Y})\ \ P(\underline{Y}) = P(\underline{Y},\underline{\theta}) = P(\underline{Y}/\underline{\theta})\ \ P(\underline{\theta})$$

وعليه فإن:

$$P(\underline{\theta}/\underline{Y}) = \frac{P(\underline{Y}/\underline{\theta})P(\underline{\theta})}{P(\underline{Y})} \quad \ldots(1)$$

إذ إن: (الجمال، ٢٠٠٣)

$$P(\underline{Y}) = \begin{cases} \int P(\underline{Y}|\underline{\theta})P(\underline{\theta})d\theta & \text{إذا كانت مستمرة} \\ \sum P(\underline{Y}|\underline{\theta})P(\underline{\theta}) & \text{إذا كانت متقطعة} \end{cases}$$

لذلك فإن :-

$$P(\underline{\theta}/\underline{Y}) \propto P(\underline{\theta}) \quad P(\underline{Y}/\underline{\theta}) \qquad \ldots(2)$$

إذ إن العلاقة $(\propto)$ تشير إلى إن الكمية تناسبية وان $\underline{\theta} = (\theta_1,...,\theta_P)$ تمثل متجه المعالم غير المعرفة ذات البعد P .

وعليه فان $P(\underline{\theta}/\underline{Y})$ تمثل دالة الكثافة الاحتمالية اللاحقة لمتجه المعالم $\underline{\theta}$ وان الصيغة (٢) تمثل المعلومات المتاحة كافة المتمثلة بمعلومات العينة والمعلومات السابقة.

التحليل التمييزي :-

يعد أسلوب التحليل التمييزي الخطي احد الأساليب الإحصائية لتحليل المتغيرات المتعددة ذات الطبيعة التنبؤية والذي يتعامل مع هياكل البيانات التي تنطوي علي وجود عدة متغيرات توضيحية ومتغير تابع نوعي ذي طبيعة ثنائية أو ذي طبيعة متعددة، التحليل التمييزي يهتم بكيفية التمييز بين مجموعتين أو أكثر من الأفراد أو المشاهدات ، فمدير البنك مثلاً قد يرغب في تصنيف أو تمييز المستدينين إلى مستدين جيد وآخر غير جيد بناء على بعض المعايير (المتغيرات) مثل مقدار القرض المعار والدخل الشهري وعمر المستدين وهكذا، (الراوي، ٢٠٠٤)

إن المشكلة الأساسية التي يتعامل معها التحليل التمييزي الخطي تتمثل في انه في بعض الأحيان قد يوجد مجتمعان إحصائيان محددان ولكل منهما صفاته المميزة له إلا أنهما يتداخلان فيما بينهما، بمعنى توجد مجموعة من المفردات التي تحمل خصائص قياسها بعض صفات المجتمع الثاني وتشكو هذه المجموعة من المفردات ما يسمي بمنطقة التداخل وهنا يتطلب الأمر إيجاد قاعدة إحصائية يتم علي أساسها التمييز بين مفردات المجتمعين بالشكل الذي يؤدي إلي تحديد منطقة التداخل، إن الهدف الرئيس من التحليل التمييزي هو إرجاع المشاهدة أو مجموعة المشاهدات إلى مجاميعها الأصلية وبأقل خطأ ممكن. (حمودات، ٢٠٠٥)

تحليل بيز التمييزي Bayesian discriminant analysis :-

في هذه الطريقة تتوفر لدينا معلومات أولية عن كل المجاميع (يأتي عن طريق التجارب السابقة أو الخبرة الشخصية) بإن المفردة المسحوبة تنتمي إلى مجموعة من المجاميع الموجودة، وتكون هذه المعلومات على شكل توزيع احتمالي ويدعى بالاحتمال المسبق (الأولي)، والتي تساعد في تقليل خطا التصنيف إلى اقل ما يمكن وبالاعتماد على الاحتمال الأولي ودالة الاحتمال الشرطية نستطيع إيجاد الاحتمال اللاحق (النهائي) كما يأتي:- (مولود، ٢٠٠٠).

$$P(A_i / B) = \frac{P(B / A_i)\,P(A_i)}{\sum_{i=1}^{n} P(B / A_i)\,P(A_i)} \qquad i = 1,2,\ldots,n \qquad \ldots(3)$$

حيث أن :-

$P(A_i)$: الاحتمال المسبق لتصنيف المشاهدة X إلى المجموعة i .

$P(B/A_i)$: دالة الاحتمال الشرطية (وهي دالة الاحتمال المشترك لـ n من المتغيرات العشوائية).

$P(A_i/B)$: الاحتمال اللاحق.

والآن لتكن لدينا المجموعتان G_1 و G_2 فان الاحتمال النهائي لسحب المشاهدة X من المجموعة الأولى G_1 يكون كما يأتي:-

$$P(G_1 / X) = \frac{P(X / G_1)\,P(G_1)}{\sum_{i=1}^{2} P(X / G_i)\,P(G_i)} \qquad \ldots(4)$$

وان الاحتمال النهائي لسحب المشاهدة X من المجموعة الثانية G2 يكون كما يأتي :-

$$P(G_2 / X) = \frac{P(X / G_2)\,P(G_2)}{\sum_{i=1}^{2} P(X / G_i)\,P(G_i)} \qquad \ldots(5)$$

جمع البيانات

أخذت البيانات من المعهد التقني /نينوى (اللجنة الامتحانية) وتشمل عينة الدراسة على (٨٥) طالباً وطالبة وتمثل هذه البيانات درجات هؤلاء الطلبة للدورين الأول والثاني ولجميع المواد الدراسية في قسم المحاسبة، ويمكن توضيح المتغيرات بما يأتي:-

X_1 : مادة مبادئ محاسبة .

X_2 : مادة تطبيقات حاسبة .

X_3 : مادة مبادئ إحصاء .

X_4 : مادة حقوق الإنسان .

X_5 : مادة الاقتصاد والمالية العامة .

X_6 : مادة محاسبة التكاليف .

X_7 : مادة المحاسبة الحكومية .

X_8 : مادة مبادئ الإدارة .

Y : تمثل نوع الدراسة (1 = صباحي، 2= مسائي)

التحليل الإحصائي

استخدم البرنامج الإحصائي (Spss V12) لإيجاد تحليل بيز التمييزي من خلال إيجاد الاحتمالات الأولية لكلتا المجموعتين فضلاً عن إيجاد التوزيع السابق واللاحق للمجموعة الأولى والتوزيع اللاحق للمجموعة الثانية فضلاً عن ذلك تم إيجاد مربع المسافة الاقليدية لكلتا المجموعتين فضلاً عن دالة التمييز والنتائج موضحة أدناه :-

جدول (١): يبين الاحتمال الأولي لكلتا المجموعتين (الصباحية والمسائية)

نوع الدراسة	الاحتمال الأولي	الحالات المستخدمة في التحليل
صباحي	0.765	65
مسائي	0.235	20
الكلي	1	85

جدول (٢): يمثل نتائج التصنيف لكلتا المجموعتين (الصباحية والمسائية)

نوع الدراسة	صباحية	مسائية	الكلي
صباحية	60	5	65
مسائية	12	8	20
صباحية (%)	92.3	7.7	100.0
مسائية (%)	60.0	40.0	100.0

من الجدول السابق يتضح لنا إن هنالك (٦٠) مشاهدة من المجموعة الأولى (الدراسة الصباحية) تم تصنيفها بصورة صحيحة وان هنالك (٥) مشاهدات تم تصنيفها بصورة خاطئة وان نسبة خطأ التصنيف للمجموعة الأولى بلغت (7.7%)، أما بالنسبة للمجموعة الثانية (الدراسة المسائية) تبين أن هنالك (٨) مشاهدات تم تصنيفها بصورة صحيحة وان هنالك (١٢) مشاهدة وان نسبة خطأ التصنيف بلغت (60%) لذلك يمكن تفسير ارتفاع نسبة خطأ التصنيف في المجموعة الثانية (الدراسة المسائية) إلى قلة عدد الطلبة والبالغ عددهم (٢٠) قياساً بعدد متغيرات الدراسة والبالغ عددها (٨).

جدول (٣): يبين التوزيعات السابقة واللاحقة لكلتا المجموعتين فضلاً عن مربع المسافة الاقليدية ودالة التمييز

الحالة		المجموعة الأولى				المجموعة الثانية			دالة التمييز
	المجاميع بصورة عامة	المجموعة التي من المفترض أن ينتمي إليها الطالب	$P(X/G_1)$	$P(G_1/X)$	مربع المسافة	المجموعة	$P(G_2/X)$	مربع المسافة	
1	1	1	.844	.868	.039	2	.132	1.442	.133
2	1	**2****	.461	.696	.544	1	.304	4.557	-1.806
3	1	1	.170	.560	1.879	2	.440	.001	-1.042
4	1	1	.180	.983	1.798	2	.017	7.498	1.670
5	1	1	.534	.783	.387	2	.217	.600	-.294
6	1	1	.561	.951	.338	2	.049	3.916	.910
7	1	1	.326	.686	.966	2	.314	.172	-.654
8	1	1	.148	.985	2.096	2	.015	8.093	1.776
9	1	1	.392	.966	.733	2	.034	5.079	1.185
10	1	1	.959	.889	.003	2	.111	1.812	.278
11	1	1	.406	.965	.691	2	.035	4.967	1.160
12	1	1	.313	.678	1.016	2	.322	.152	-.679
13	1	1	.533	.783	.388	2	.217	.600	-.294
14	1	1	.045	.993	4.029	2	.007	11.590	2.336
15	1	1	.782	.927	.076	2	.073	2.801	.605
16	1	1	.469	.960	.524	2	.040	4.499	1.053
17	1	1	.926	.908	.009	2	.092	2.221	.422
18	1	1	.946	.887	.005	2	.113	1.768	.261
19	1	1	.326	.971	.966	2	.029	5.664	1.311
20	1	1	.510	.775	.433	2	.225	.546	-.329
21	1	1	.850	.869	.036	2	.131	1.459	.139
22	1	1	.146	.985	2.117	2	.015	8.136	1.784
23	1	1	.203	.593	1.619	2	.407	.016	-.944
24	1	1	.241	.978	1.373	2	.022	6.599	1.500
25	1	1	.695	.937	.153	2	.063	3.200	.720
26	1	1	.976	.892	.001	2	.108	1.868	.298
27	1	1	.036	.994	4.392	2	.006	12.201	2.425
28	1	1	.345	.970	.890	2	.030	5.480	1.272
29	1	**2****	.874	.505	.025	1	.495	2.422	-1.228
30	1	1	.765	.929	.089	2	.071	2.875	.627
31	1	1	.969	.891	.002	2	.109	1.845	.290
32	1	1	.565	.951	.332	2	.049	3.894	.905
33	1	1	.456	.961	.556	2	.039	4.591	1.074
34	1	1	.761	.849	.093	2	.151	1.193	.024
35	1	1	.454	.752	.560	2	.248	.421	-.420
36	1	1	.553	.952	.352	2	.048	3.963	.922
37	1	1	.347	.698	.886	2	.302	.208	-.612
38	1	1	.293	.974	1.104	2	.026	5.993	1.379
39	1	1	.457	.753	.553	2	.247	.427	-.415
40	1	1	.015	.996	5.930	2	.004	14.687	2.764

.466	2.354	.087	2	.019	.913	.891	1	1	41
.275	1.804	.111	2	.003	.889	.957	1	1	42
1.090	4.658	.038	2	.579	.962	.447	1	1	43
-.003	1.135	.156	2	.110	.844	.740	1	1	44
.023	1.191	.151	2	.094	.849	.760	1	1	45
1.716	7.752	.016	2	1.924	.984	.165	1	1	46
1.600	7.120	.019	2	1.616	.981	.204	1	1	47
.426	2.233	.092	2	.009	.908	.923	1	1	48
-.206	.744	.197	2	.286	.803	.593	1	1	49
.520	2.522	.082	2	.036	.918	.849	1	1	50
-.274	.631	.212	2	.363	.788	.547	1	1	51
-1.295	2.636	.472	1	.051	.528	.821	**2****	1	52
.482	2.405	.086	2	.024	.914	.878	1	1	53
-.857	.045	.378	2	1.407	.622	.236	1	1	54
.039	1.226	.148	2	.084	.852	.772	1	1	55
.876	3.782	.051	2	.300	.949	.584	1	1	56
.133	1.444	.132	2	.038	.868	.845	1	1	57
-.696	.139	.327	2	1.050	.673	.305	1	1	58
-.114	.911	.177	2	.196	.823	.658	1	1	59
-.164	.817	.188	2	.243	.812	.622	1	1	60
.477	2.387	.086	2	.022	.914	.883	1	1	61
-2.669	8.986	.116	1	2.561	.884	.110	**2****	1	62
-1.949	5.187	.264	1	.775	.736	.379	**2****	1	63
.009	1.160	.153	2	.102	.847	.749	1	1	64
.958	4.107	.046	2	.396	.954	.529	1	1	65
-1.177	.012	.487	2	2.268	.513	.132	**1****	2	66
-1.309	2.681	.467	1	.058	.533	.810	2	2	67
-1.104	.001	.462	2	2.054	.538	.152	**1****	2	68
-1.190	.015	.492	2	2.307	.508	.129	**1****	2	69
-2.600	8.576	.126	1	2.345	.874	.126	2	2	70
-.298	.594	.218	2	.392	.782	.531	**1****	2	71
-1.101	.001	.461	2	2.044	.539	.153	**1****	2	72
.400	2.157	.095	2	.005	.905	.943	**1****	2	73
-1.030	.002	.436	2	1.845	.564	.174	**1****	2	74
-1.675	4.017	.344	1	.368	.656	.544	2	2	75
-.448	.385	.256	2	.604	.744	.437	**1****	2	76
-2.200	6.397	.201	1	1.281	.799	.258	2	2	77
-1.250	2.492	.487	1	.033	.513	.856	2	2	78
-1.378	2.912	.443	1	.096	.557	.757	2	2	79
-1.557	3.555	.382	1	.238	.618	.625	2	2	80
-.042	1.053	.163	2	.138	.837	.711	**1****	2	81
.283	1.826	.110	2	.002	.890	.963	**1****	2	82
-.639	.185	.309	2	.936	.691	.333	**1****	2	83
-.572	.247	.290	2	.811	.710	.368	**1****	2	84
-2.483	7.906	.145	1	2.001	.855	.157	2	2	85

(**): تدل على خطأ في التصنيف.

الجدول السابق يوضح المشاهدات التي ظهر بها خطأ التصنيف كما يوضح التوزيع السابق واللاحق لكلتا المجموعتين.

الاستنتاجات والتوصيات

١) يمكن الاعتماد على المعادلتين (٤) و (٥) لإيجاد التوزيع اللاحق لكل من المجموعتين ومن ثم نحصل على تحليل بيز التمييزي.

٢) أظهرت النتائج أن هنالك (٥) مشاهدات من بين (٦٥) مشاهدة تم تـصنيفها بـصورة خاطئة بالنسبة للمجموعة الأولى (الدراسة الصباحية).

٣) أظهرت نتائج المجموعة الثانية (الدراسة المسائية) أن هنالك (١٢) مشاهدة مـن بـين (٢٠) مشاهدة تم تصنيفها بصورة خاطئة.

٤) تبين من النتائج أن نسبة خطأ التصنيف للمجموعة الأولى (الدراسة الصباحية) والبـالغ (7.7%) اقل بكثير من نسبة خطأ التصنيف للمجموعة الثانية (الدراسة المسائية) والبالغ (60%).

المصادر

١) الجمال، زكريا يحيى نوري (2003): "**تقدير بيز الهرمي الحركي مع المحاكاة**"، رسـالة ماجستير، كلية علوم الحاسبات والرياضيات، جامعة الموصل.

٢) الراوي، عمر فوزي صالح بدر (2004) :"**استخدام الدالة التمييزية في السيطرة النوعية مع تطبيق على ولادات الأطفال الخـدج**"، رسـالة ماجـستير، كليـة علـوم الحاسـبات والرياضيات، جامعة الموصل.

٣) الرجبو، عمار حازم طه (٢٠٠٥): "**بعض طرائق المقدرات التقليدية ومقدر بيز لمعلمـات نموذج الانحدار الخطي العام (دراسة مقارنة مع تطبيق في مجال طبي)**"، رسالة ماجستير، كلية علوم الحاسبات والرياضيات، جامعة الموصل.

٤) حمودات، آلاء عبد الستار داؤد (٢٠٠٥): "**الدالة التمييزية وطـرق تحديـد متغيراتهـا**"، رسالة ماجستير، كلية علوم الحاسبات والرياضيات، جامعة الموصل.

٥) مولود، كردستان إبراهيم (2000) :"**استخدام التحليل المميـز لتـشخيص أهـم العوامـل المؤثرة في التصنيف السريري لأمراض القلب**"، رسالة ماجستير، كلية الإدارة والاقتصاد، جامعة صلاح الدين.