



ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Instituto de Ciencias Matemáticas

INGENIERÍA EN ESTADÍSTICA INFORMÁTICA

“Diseño e implementación de un aplicativo web para el aprendizaje de análisis discriminante”

TESIS DE GRADO

Previa la obtención del título de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

Presentada por:

César E. Noboa Cisneros



GUAYAQUIL – ECUADOR

AÑO

2008

TRIBUNAL DE GRADUACIÓN

Mat. John Ramírez
PRESIDENTE

Ing. Juan Alvarado
DIRECTOR DE TESIS

Ing. Félix Ramírez
VOCAL

Ing. Pablo Álvarez
VOCAL

DECLARACIÓN EXPRESA

“La responsabilidad del contenido de esta Tesis de Grado, me corresponde exclusivamente; y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL”

César Noboa Cisneros

RESUMEN

El objetivo principal de esta tesis es desarrollar una página web para que cualquier estudiante con acceso a Internet la utilice como herramienta gráfica de apoyo para el aprendizaje de análisis discriminante. Con esta página web el estudiante podrá experimentar con 5 métodos de discriminación aplicados a distintas muestras de datos teniendo total libertad para modificar o formar su propia muestra, todo esto visualmente. También podrá hacer comparaciones entre las curvas discriminantes de distintos métodos.

La presente tesis se inicia exponiendo la tarea de la discriminación y la Minería de Datos como disciplina útil para abordar esta tarea. Luego se describen con cierto nivel de detalle los métodos que estarán disponibles al usuario en la página web.

En el segundo capítulo se expone brevemente la principal herramienta informática utilizada para el desarrollo de esta aplicación web: JavaScript. En términos generales, en este capítulo se describe una parte de los comandos e instrucciones de JavaScript que atañen a este trabajo. Ya finalizando el capítulo también se hace referencia a la librería `wz_dragdrop`,

útil para convertir imágenes fijas en movibles y a la librería `wz_jsgraphics`, útil para dibujar figuras geométricas en una página web.

El tercer capítulo presenta el diseño de la página web. En este capítulo se encuentran descritas las diversas opciones y posibilidades que ofrece la página junto con algunas explicaciones adicionales.

Aunque el usuario puede tener su propia experiencia con la página web, el último capítulo está escrito a manera de ejemplo de esta posible experiencia. Aquí se reflejan los resultados de una serie de ensayos que aplican los 5 métodos descritos en el capítulo 1 a diferentes conjuntos de datos. A la luz de los resultados obtenidos en la página se ejemplifican algunas características de cada uno de los métodos y se efectúan ciertas comparaciones entre algunos de ellos. Puesto que esta página es una herramienta especialmente gráfica la visualización de las distintas curvas discriminantes tendrá un papel sobresaliente en este capítulo.

Al final de esta tesis se presentan algunas conclusiones relacionadas a las pruebas realizadas con la página y algunas recomendaciones que podrían ayudar a aquellos interesados en el tema a tener una experiencia de aprendizaje satisfactoria con la página web.

ÍNDICE GENERAL

	Pág.
RESUMEN.....	II
ÍNDICE GENERAL.....	III
ÍNDICE DE FIGURAS.....	IV
ÍNDICE DE TABLAS.....	V
ÍNDICE DE CUADROS.....	VI
INTRODUCCIÓN.....	1
1. ANÁLISIS DISCRIMINANTE	3
1.1 Minería de Datos	8
1.2 Método de los k-vecinos más cercanos.....	14
1.3 Método de Naive Bayes Kernel.....	16
1.4 Método de regresión lineal	20
1.5 Método de regresión logística	24
1.6 Árbol de decisión	28

2. JAVASCRIPT (JSCRIPT)	35
2.1 Introducción	35
2.2 Referencia rápida.....	39
2.3 Librería JavaScript Vector Graphics: wz_jsgraphics.js	58
2.4 Librería JavaScript Drag & Drop: wz_dragdrop.js.....	63
3. DISEÑO DEL APLICATIVO WEB	65
3.1 Distribución del conjunto de datos.....	67
3.2 Selección de parámetros.....	72
3.3 Botones de opción.....	74
3.4 Tablero de resultados.....	78
4. RESULTADOS DE LA PÁGINA	80
4.1 Método de los k-vecinos más cercanos.....	81
4.2 Método de Naive Bayes Kernel.....	89
4.3 Método de árbol de decisión.....	97
4.4 Métodos de regresión.....	109
4.4.1 Regresión lineal y regresión logística	114
4.5 Experimentos de comparación entre métodos	119

CONCLUSIONES Y RECOMENDACIONES

BIBLIOGRAFÍA

ÍNDICE DE FIGURAS

	Pág.
Figura 1.1	ILUSTRACIÓN DE UNA CURVA DISCRIMINANTE..... 6
Figura 1.2	ERROR DE ENTRENAMIENTO..... 10
Figura 1.3	GRADOS DE EXPRESIVIDAD..... 11
Figura 1.4	ILUSTRACIÓN DE LOS 5-VECINOS MÁS CERCANOS.. 15
Figura 1.5	LIMITACIÓN DE K-VECINOS PARA K=1..... 16
Figura 1.6	ESQUEMA DE LA DISCRIMINACIÓN UTILIZANDO UN ARBOL DE DECISIÓN..... 29
Figura 1.7	RESULTADO DE UNA DISCRIMINACIÓN UTILIZANDO UN ÁRBOL DE DECISIÓN..... 31
Figura 1.8	ESQUEMA DE UN ÁRBOL DE DECISIÓN CON ATRIBUTOS NUMÉRICOS..... 32
Figura 2.1	LECTURA DE UN OBJETO ARRAY()..... 47
Figura 2.2	CÓDIGO PARA LA CREACIÓN FLEXIBLE DE BOTONES A TRAVES DEL METODO WRITE()..... 49
Figura 2.3	CREACIÓN FLEXIBLE DE BOTONES CON EL METODO WRITE()..... 50
Figura 2.4	CÓDIGO JSCRIPT PARA LA CREACIÓN DE UN ARREGLO MULTIDIMENSIONAL..... 52
Figura 2.5	PRESENTACIÓN DE UNA MATRIZ DE 4 X 20..... 53
Figura 2.6	PRESENTACIÓN DE LA PROPIEDAD <i>VALUE</i> DE UN BOTÓN..... 55
Figura 2.7	PRESENTACIÓN DE NÚMEROS UTILIZANDO EL MÉTODO RANDOM()..... 58
Figura 2.8	ESQUEMA DE GRAFICACIÓN DE UNA ELIPSE USANDO LA LIBRERÍA WZ_JSGRAPHICS.JS..... 61
Figura 2.9	VISUALIZACIÓN DE IMÁGENES COMO PUNTOS EN EL PLANO UTILIZANDO EL MÉTODO DRAWIMAGE() DE LA LIBRERÍA WZ_JSGRAPHICS.JS..... 62
Figura 3.1	PANTALLA INICIAL DEL APLICATIVO WEB..... 67

Figura 3.2	SELECCIÓN DEL MÉTODO.....	67
Figura 3.3	OPCIONES DE DISTRIBUCIÓN DE PUNTOS.....	68
Figura 3.4	DISTRIBUCIÓN DE DIAGRAMA DE VENN.....	68
Figura 3.5	CONJUNTO DE PUNTOS DISTRIBUIDOS COMO DIAGRAMA DE VENN.....	69
Figura 3.6	DISTRIBUCION DE CUADRADOS TRASLAPADOS.....	70
Figura 3.7	DISTRIBUCIÓN DE PUNTOS TIPO “EXPONENCIAL”....	71
Figura 3.8	CONJUNTO DE DATOS PERSONALIZADO.....	72
Figura 3.9	SELECCIÓN DEL NÚMERO DE VECINOS.....	73
Figura 3.10	SELECCIÓN DEL PARÁMETRO DE PARADA DE UN ÁRBOL DE DECISIÓN.....	74
Figura 3.11	BOTONES DE OPCIÓN.....	75
Figura 3.12	GRAFICACIÓN DE VARIAS CURVAS DISCRIMINANTES EN EL MISMO PLANO.....	75
Figura 3.13	DISCRIMINACION SIN VISUALIZAR DATOS DE PRUEBA.....	77
Figura 3.14	DISCRIMINACION VISUALIZANDO DATOS DE PRUEBA.....	77
Figura 3.15	VISUALIZACIÓN DEL MODELO DE UN MÉTODO.....	78
Figura 3.16	REGISTRO DE RESPUESTAS EN EL TABLERO DE RESULTADOS.....	79
Figura 4.1	MUESTRA DE LA EXPRESIVIDAD DEL MÉTODO 1-VECINO MÁS CERCANO.....	81
Figura 4.2	MUESTRA DE LA EXPRESIVIDAD DEL MÉTODO 5-VECINOS MÁS CERCANOS.....	82
Figura 4.3	CAMBIOS EN LA EXPRESIVIDAD DE LOS K-VECINOS PARA K=1, 7 Y 15.....	83
Figura 4.4	MÉTODO DE LOS K-VECINOS PARA K=1,5,10 Y 15 PARA CLASES DISTRIBUIDAS EXPONENCIALMENTE.....	85
Figura 4.5	EJEMPLO DE LA SENSIBILIDAD DE 1-VECINO MÁS CERCANO A UN DATO ANÓMALA.....	87
Figura 4.6	EJEMPLO DE LA SENSIBILIDAD DE 5-VECINOS MÁS CERCANOS A UN DATO ANÓMALA.....	87
Figura 4.7	NAIVE BAYES APLICADO A MUESTRAS DE POBLACIONES EXPONENCIALES.....	91
Figura 4.8	COMPARACION ENTRE EL MÉTODO DE NAIVE BAYES Y REGRESIÓN PARA DISTRIBUCION DE CUADRADOS TRASLAPADOS.....	92
Figura 4.9	COMPARACIÓN ENTRE 6-VECINOS Y NAIVE BAYES	94
Figura 4.10	SENSIBILIDAD DE NAIVE BAYES KERNEL ANTE UN DATO ANÓMALA, DISTRIBUCIÓN DE CUADRADOS TRASLAPADOS.....	96

Figura 4.11	SENSIBILIDAD DE NAIVE BAYES KERNEL ANTE UN DATO ANÓMALA, DISTRIBUCIÓN EXPONENCIAL.....	97
Figura 4.12	APLICACIÓN DEL ÁRBOL DE DECISIÓN VARIANDO S PARA LA MISMA MUESTRA.....	99
Figura 4.13	APLICACIÓN DE ÁRBOL DE DECISIÓN A LA DISTRIBUCIÓN DE DIAGRAMA DE VENN.....	101
Figura 4.14	APLICACIÓN DE ÁRBOL DE DECISIÓN A LA DISTRIBUCIÓN DE CUADRADOS TRASLAPADOS.....	102
Figura 4.15	VARIACIÓN DEL ARBOL DE DECISIÓN A DIFERENTES MUESTRAS CON S=2%.....	102
Figura 4.16	VARIACIÓN DEL ÁRBOL DE DECISIÓN A DIFERENTES MUESTRAS CON S=10%.....	103
Figura 4.17	ÁRBOL DE DECISIÓN PARA 2 MUESTRAS DE POBLACIONES EXPONENCIALES CON S=6%.....	104
Figura 4.18	REGRESIÓN LINEAL DE GRADO 1 Y ÁRBOL DE DECISIÓN: DIFERENCIAS AL CAPTAR EL PATRÓN DE DATOS.....	107
Figura 4.19	MODELO DE UN ÁRBOL DE DECISIÓN.....	108
Figura 4.20	REPRESENTACIÓN GRÁFICA DEL SISTEMA DE REGLAS DE LA FIGURA 4.19.....	108
Figura 4.21	REGRESION GRADO 1 APLICADO A UN CONJUNTO DE DATOS EXPONENCIAL.....	109
Figura 4.22	SENSIBILIDAD DE LA REGRESIÓN LINEAL AL CAMBIAR LA MUESTRA DE UNA POBLACIÓN CON CLASES EXPONENCIALES.....	111
Figura 4.23	REGRESION GRADO 1,2 Y 3 PARA DISCRIMINAR CLASES CON DISTRIBUCIÓN EXPONENCIAL.....	112
Figura 4.24	DIFERENCIAS EN LA EXPRESIVIDAD DE LA REGRESION DE GRADO 1 Y 2.....	113
Figura 4.25	ESBOZO DE LA SENSIBILIDAD DE LA REGRESIÓN LINEAL ANTE OUTLIERS.....	115
Figura 4.26	ESBOZO DE LA ROBUSTEZ DE LA REGRESIÓN LOGÍSTICA ANTE OUTLIERS.....	117
Figura 4.27	EJEMPLO DE UNA DIFERENCIA ENTRE REGRESIÓN LINEAL Y REGRESIÓN LOGÍSTICA.....	118
Figura 4.28	MÉTODOS DE REGRESIÓN GRADO=2 APLICADOS A UN CONJUNTO CON CLASES DISTRIBUIDAS EXPONENCIALMENTE.....	119
Figura 4.29	DISTRIBUCIÓN DE CÍRCULOS CONCÉNTRICOS.....	122
Figura 4.30	REGRESIÓN DE GRADO 1 Y 2 APLICADOS A LA DISTRIBUCIÓN DE CÍRCULOS CONCÉNTRICOS.....	124

Figura 4.31	EXPRESIVIDAD DE LA REGRESION POLINOMIAL GRADO 4 APLICADA A LA DISTRIBUCIÓN DE CÍRCULOS CONCÉNTRICOS.....	125
Figura 4.32	ÁRBOL DE DECISIÓN APLICADO A LA DISTRIBUCIÓN DE CÍRCULOS CONCÉNTRICOS.....	126

ÍNDICE DE TABLAS

		Pág.
Tabla I	EJEMPLO DE REGISTROS PARA UN ANÁLISIS DISCRIMINANTE.....	4
Tabla II	ERRORES DEL MÉTODO K-VECINOS PARA K=1, 7 Y 15 APLICADOS A LA DISTRIBUCIÓN DE CUADRADOS TRASLAPADOS.....	84
Tabla III	ERROR DE PRUEBA DE 1-VECINO VS. 5-VECINOS PARA MUESTRAS CON CLASES EXPONENCIALES..	88
Tabla IV	RESULTADOS DEL ÁRBOL DE DECISIÓN EN LA DISTRIBUCIÓN DE DIAGRAMA DE VENN.....	100
Tabla V	RESULTADOS DEL ÁRBOL DE DECISIÓN EN LA DISTRIBUCIÓN DE CUADRADOS TRASLAPADOS....	100
Tabla VI	ÁRBOL DE DECISIÓN APLICADO A 6 MUESTRAS DE LA MISMA POBLACION CON CLASES DISTRIBUIDAS EXPONENCIALMENTE.....	105
Tabla VII	APLICACIÓN DE REGRESION GRADO 1 A 10 MUESTRAS PROVENIENTES DE UNA POBLACIÓN CON CLASES DISTRIBUIDAS EXPONENCIALMENTE	111
Tabla VIII	RESULTADOS DE UN EXPERIMENTO CON CLASES DISTRIBUIDAS EXPONENCIALMENTE.....	120
Tabla IX	RESULTADOS DE UN EXPERIMENTO CON CLASES DISTRIBUIDAS EN CÍRCULOS CONCÉNTRICOS.....	125

ÍNDICE DE CUADROS

	Pág.
Cuadro 1.1 ALGUNAS FUNCIONES NÚCLEO $K(x)$ UTILIZADAS	19