Introducción al reconocimiento de patrones: Trabajo práctico 1, repaso de Probabilidad.

M. Sc. Saúl Calderón Ramírez
Instituto Tecnológico de Costa Rica,
Escuela de Computación, bachillerato en Ingeniería en Computación,
PAttern Recongition and MAchine Learning Group (PARMA-Group)

7 de marzo de 2017

Fecha de entrega: 5 de Marzo del 2017.
Entrega: Un archivo .zip con el código fuente LaTeX o Lyx, el pdf, y un script en MATLAB, debidamente documentado, con una función definida por ejercicio. A través del TEC-digital.
Modo de trabajo: Grupos de 3 personas.

Resumen

En el presente trabajo práctico se repasarán aspectos básicos de las probabilidades, relacionados con los conceptos a desarrollar a lo largo del curso, mezclando aspectos teóricos y prácticos, usando el lenguaje MATLAB.

1. Funciones de densidad de probabilidad (35 puntos)

1. Supongamos que tenemos tres cajas coloreadas r (rojo), a (azul) y v (verde). La caja r contiene 3 manzanas, 4 naranjas, y 3 limones, la caja a contiene 1 manzana, 1 naranja, y 0 limones, y la caja v contiene 3 manzanas, 3 naranjas y 4 limones. Si se selecciona una caja al azar con probabilidades $p(r) = 0,2$, $p(a) = 0,2$, $p(v) = 0,6$, y una fruta se saca de la caja (con la misma probabilidad de seleccionar cualquiera de las frutas en la caja), entonces:

   a) ¿Cuál es la probabilidad de seleccionar una manzana? ¿Esto corresponde a una probabilidad a priori o a posteriori?

   b) Si observamos que la fruta seleccionada es en realidad una naranja, ¿cuál es la probabilidad de que venga de la caja verde? ¿Esto corresponde a una probabilidad a priori o a posteriori?

2. Demuestre que la esperanza $\mathbb{E}[X]$ de una variable aleatoria X descrita por una distribución normal $\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$ está dada por μ.

3. Demuestre que la moda (el máximo), de una función Gaussiana $f(x)$, con $f : \mathbb{R} \to \mathbb{R}$ está dada por μ.

4. Demuestre que la moda (el máximo), de una función Gaussiana $f(\vec{x})$, con $f : \mathbb{R}^n \to \mathbb{R}$ está dada por $\vec{\mu} \in \mathbb{R}^n$.

5. Utilizando las funciones *imshow* y *surf* de MATLAB (en cada figura incluya las dos gráficas), grafique, de ser posible, las siguientes funciones Gaussianas $f : \mathbb{R}^2 \to \mathbb{R}$, con los siguientes parámetros. Programe la función manualmente usando la definición de la función Gaussiana multivariable:

   a) $\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 10 \end{bmatrix}$, $\vec{\mu} = \begin{bmatrix} 20 \\ 30 \end{bmatrix}$

   b) $\Sigma = \begin{bmatrix} 10 & 0 \\ 0 & 2 \end{bmatrix}$, $\vec{\mu} = \begin{bmatrix} 30 \\ 20 \end{bmatrix}$

   c) $\Sigma = \begin{bmatrix} 10 & -1{,}5 \\ -1{,}5 & 2 \end{bmatrix}$, $\vec{\mu} = \begin{bmatrix} 15 \\ 20 \end{bmatrix}$

   d) $\Sigma = \begin{bmatrix} 10 & 20 \\ 20 & 2 \end{bmatrix}$, $\vec{\mu} = \begin{bmatrix} 20 \\ 15 \end{bmatrix}$

   e) $\Sigma = \begin{bmatrix} 10 & 20 \\ 25 & 2 \end{bmatrix}$, $\vec{\mu} = \begin{bmatrix} 20 \\ 15 \end{bmatrix}$

   y comente los efectos que ejerce **cada** desviación estándar $\Sigma_{1,1} = \sigma_1$ y $\Sigma_{2,2} = \sigma_2$, además de **cada** covarianza $\Sigma_{1,2}$ y $\Sigma_{2,1}$ en la forma de la función de densidad Gaussiana.

6. Para una matriz de datos A programe un método que calcule su matriz de covarianza, usando funciones como *mean* y *repmat*.

7. Genere dos conjuntos de datos con $N = 100$ puntos en un espacio $\mathbb{R}^2$ a partir de una distribución Gaussiana, donde el primer cúmulo de datos se tenga que $\Sigma_1 = \begin{bmatrix} 5 & 0 \\ 0 & 2 \end{bmatrix}$ y para el segundo cúmulo de datos $\Sigma_1 = \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}$, usando la función *mvnrnd*, el primero centrado en $\vec{\mu}_1 = \begin{bmatrix} 0 & 0 \end{bmatrix}$ y el segundo en $\vec{\mu}_2 = \begin{bmatrix} 10 & 10 \end{bmatrix}$.

   a) Grafique los resultados usando la función *scatter*, con símbolos que dejen claro la pertenencia a cada uno de los cúmulos.

2. Introducción al análisis de componentes principales (35 puntos)

Usando MATLAB, realice un análisis de componentes principales, desarrollando los siguientes pasos:

1. Escriba la función *generarPuntos*, la cual genere $n = 20$ puntos en $\vec{x}_i \in \mathbb{R}^3$ aleatorios los cuales pertenezcan a un plano con función $f(x, y) = 0{,}2x + y + \epsilon$, $f : \mathbb{R}^2 \to \mathbb{R}$, con ϵ una variable aleatoria de $\mu = 0$ y $\sigma = 0{,}05$. Almacenelos en una matriz de modo que:

$$X = \begin{bmatrix} | & | & | \\ \vec{x}_1 & \dots & \vec{x}_n \\ | & | & | \end{bmatrix}$$

   a) Grafique los puntos con la función *scatter3*.

2. ¿Cuales deben ser las dimensiones de la matriz de covarianza Σ para tales datos? Calculela usando las funciones apropiadas en MATLAB.

3. Calcule los auto-vectores y auto-valores de tal matriz de covarianza Σ.

4. Tome los dos auto-vectores de Σ con mayores auto-valores $\vec{v}_1$ y $\vec{v}_2$ para crear un nuevo subespacio $E = \text{espacioGenerado}\{\vec{v}_1, \vec{v}_2\}$, y cree la matriz de la base:

$$V = \begin{bmatrix} | & | \\ \vec{v}_1 & \vec{v}_2 \\ | & | \end{bmatrix}.$$

5. Verifique si tales auto-vectores son orto-normales, si es así, ¿porqué sucede esto?

6. Calcule la muestra promedio $\vec{\mu} \in \mathbb{R}^3$ para los datos en X, y calcule una nueva matriz X' en la que cada vector en el espacio tenga su origen en $\vec{\mu}$, haciendo que cada columna i esté dada por $\vec{u}_i = \vec{x}_i - \vec{\mu}$. en $X_\mu = \begin{bmatrix} | & | & | \\ \vec{u}_1 & \dots & \vec{u}_n \\ | & | & | \end{bmatrix}$

7. Demuestre que si, en general, los vectores generadores de un subespacio $E = \text{espacioGenerado}\{\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n\}$, con $V = [\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n]$ son orto-normales, la ecuación de la proyección de un vector $\vec{u}$ originalmente dada por:

$$\text{proy}(\vec{u}; V) = \text{argmin}_{\vec{v} \in E} \|\vec{v} - \vec{u}\|_2 = V \left(V^T V\right)^{-1} V^T \vec{u},$$

   se puede reescribir como:

$$\text{proy}(\vec{u}; V) = \text{argmin}_{\vec{v} \in E} \|\vec{v} - \vec{u}\|_2 = V V^T \vec{u},$$

8. Proyecte, usando la ecuación anterior de proyección, todos los datos en la matriz X_μ y almacenelos en una nueva matriz X_p. Compruebe que la simplificación de la proyección propuesta en el apartado anterior genera los mismos resultados (use ambas fórmulas pero exprese la proyección en términos de la matriz X, proy $(X; V)$).

   a) Compruebe además que la minimización realizada al encontrar la proyección de cada uno de los vectores (X_p), logra una distancia eucliciana de cero con los vectores originales en X_μ.

9. Grafique los puntos obtenidos en la matriz X usando la función *scatter3*, y grafique en la misma figura los 2 auto-vectores que forman el espacio generador con la función *quiver3*, con origen en la media de los datos o centroide μ. Comente los resultados.

10. Reduzca la dimensionalidad de los datos, de modo que se pase de un espacio en $\mathbb{R}^3$ a un espacio en $\mathbb{R}^2$ usando 2 los auto-vectores con mayores auto-valores:

    a) Para cada muestra $\vec{u}_i$ calcule la magnitud de la proyección en cada eje del nuevo espacio vectorial $E_1 = \text{espacioGenerado}\{\vec{v}_1, \vec{v}_2\}$, creando una muestra con dimensión reducida $\vec{x'}_i = \begin{bmatrix} x_{i,1} \\ x_{i,2} \end{bmatrix}$ donde:

$$x_{i,1} = \vec{u}_i \cdot \vec{v}_1$$
$$x_{i,2} = \vec{u}_i \cdot \vec{v}_2$$

    y comente ¿Es la formulación anterior para la proyección en cada eje $\vec{v}_1$ y $\vec{v}_2$ equivalente a la fórmula $\frac{\vec{u}_i \cdot \vec{v}_j}{\|\vec{v}_j\|}$? ¿Porqué?

    1) Agrupe los resultados en la matriz $X' = \begin{bmatrix} | & & | \\ \vec{x'}_1 & \dots & \vec{x'}_n \\ | & & | \end{bmatrix}$ y grafiquelos usando la función *scatter2*. Comente los resultados, ¿Realmente hubo una reducción de la dimensionalidad, y se preservarvaron los ejes de mayor varianza?

3. Ruido aditivo Gaussiano (30 puntos)

1. Abrir las imágenes provistas utilizando las funciones *imread* de MATLAB.

2. Convertir las imágenes en colores a escala de grises.

3. Implementar una función que contamine imágenes con ruido gaussiano e impulsivo, utilizando la función de la biblioteca de MATLAB *imnoise*. Use el comando help para conocer sus parámetros, y escojalos de manera que la tasa de ruido razonable.

   a) Escoja dos conjuntos de parámetros distintos y documentelos.
   b) Guarde las imágenes en el disco duro, usando para ello la función *imwrite.*

4. Implementar una función *calcularHist* que itere sobre todos los píxeles de una imagen en escala de grises determinada, y que cuente las ocurrencias de cada valor de intensidad, retornando un arreglo con la cantidad de ocurrencias total de todos los valores en la escala de grises (0 a 255).

   a) Normalice el valor de cada casilla por el número total de pixeles en la imagen.

5. En las imágenes de prueba de escala de grises contaminadas con ruido previamente, recortar al menos 3 regiones uniformes y guardarlas en un archivo nuevo.

   a) Por cada imagen, calcular y graficar el histograma, usando la función *plot* y rotulando correctamente los ejes de la gráfica.
   b) Comente los resultados obtenidos respecto a lo visto en clase.