Extracción de características en imágenes

Uso de Procesos Gaussianos para clasificación

Autor:
Nicolás Delgado Guerrero

Máster:
Ciencia de Datos e Ingeniería de Computadores

ÍNDICE

Índice de tablas

Índice de figuras

Práctica III: Uso de Procesos Gaussianos para clasificación.

El objetivo de esta práctica es aprender a utilizar los Procesos Gaussianos (GP) en un problema de clasificación y discutir los resultados obtenidos. Vamos a utilizar una base de datos de imágenes histológicas de cáncer de próstata. Para abordar el problema de clasificación se han utilizado bloques. Un problema importante es decidir su tamaño. Cada muestra o instancia corresponde a un bloque de tamaño 2048x2048 y ha sido clasificado por un anatomopatólogo como cancerígeno o no cancerígeno.

Para cada bloque 2048x2048 hemos calculado el histograma de los rasgos obtenidos tras calcular características LBP uniformes invariantes por rotaciones con radio 1 y número de vecinos igual a 8 sobre cada uno de los píxeles del parche.

Se proporciona un total de 1312 instancias, cada una de ellas corresponde a un bloque de tejido histopatológico de próstata con tamaño 2048x2048. De estas instancias, 1014 proceden de tejido sano y el resto, 298, de tejido cancerígeno.

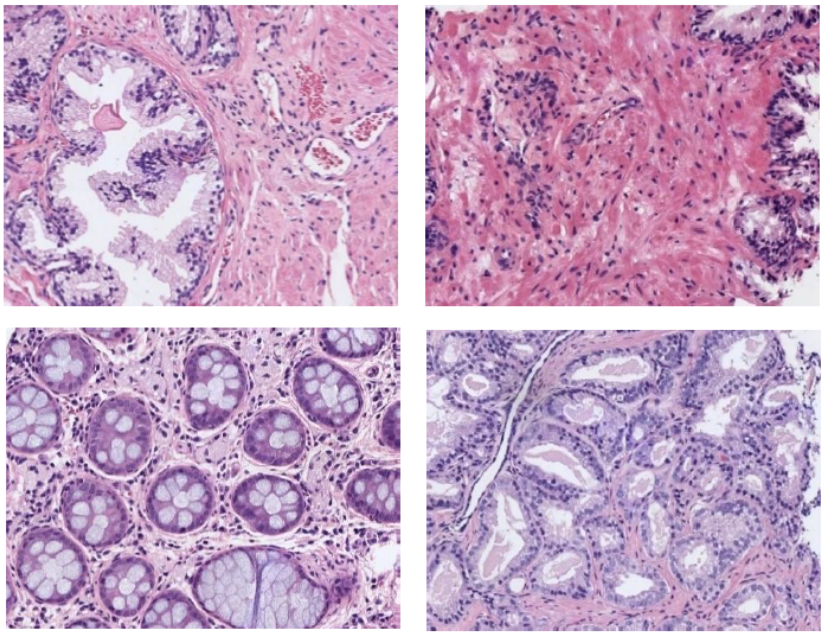

Figura 1: Imágenes de tejido sano (panel superior) y de tejido cancerígeno (panel inferior).

¿Qué es un proceso Gaussiano?

Un proceso Gaussiano es un caso particular de lo que en teoría de la probabilidad se conoce como proceso estocástico. Definiremos que es un proceso estocástico y que particularidades debe verificar para que este se denomine proceso Gaussiano.

Definición 1 *(Proceso Estocástico)*
Se define un proceso estocástico como una colección de variables aleatorias definidas sobre un mismo espacio de probabilidad $(\Omega, \mathcal{A}, P)$, donde Ω es el espacio de sucesos, $\mathcal{A}$ una $\sigma-$álgebra y P una medida de probabilidad. Las variables aleatorias están indexadas por algún conjunto T; además todas ellas toman valores en un mismo espacio S que debe ser medible respecto alguna $\sigma-$álgebra $\tilde{\mathcal{A}}$.

En otras palabras, para un espacio de probabilidad dado $(\Omega, \mathcal{A}, P)$ y un espacio de medida $(S, \tilde{\mathcal{A}})$, un proceso estocástico es una colección de variables aleatorias valuadas en S que pueden representarse como:

$$\{X(t, \omega) \mid t \in T, \omega \in \Omega\}.$$

Veamos que condición adicional debe de cumplir la colección de variables aleatorias de un proceso estocástico para que este sea Gaussiano.

Definición 2 *(Proceso Gaussiano)*
Un proceso estocástico $\{X(t, \omega) \mid t \in T, \omega \in \Omega\}$, se dice que es Gausiano si para cualquier subconjunto finito $\{X(n, \omega) \mid n \in \{1, \ldots, N\} \subset T, \omega \in \Omega\}$, este sigue una distribución normal.

Dado un proceso Gaussiano $\mathbf{Y}$ si tomamos un subconjunto finito $\{Y_1, Y_2, \ldots, Y_N\} \subset \mathbf{Y}$, este va seguir una distribución normal multivariante $(Y_1, Y_2, \ldots, Y_N)^T \to \mathcal{N}_N(\mu, \Sigma)$ donde $\mu \in \mathbb{R}^N$ es un vector de medias y $\Sigma \in M_n(\mathbb{R})$ una matriz de covarianzas.

Conectando con nuestro problema de clasificación; el estado del mundo viene dado por $\mathbf{w} = \{1, -1\}$. Tendremos 1 cuando nuestra imagen presente cáncer y -1 en caso contrario. Queremos conocer la distribución de que una imagen presente cáncer dada una nueva imagen. Para ello debemos definir una distribución a priori sobre las imágenes y un modelo de observación sobre el estado del mundo dada las imágenes. Si:

$$\mathbf{w} = f(\mathbf{X}) + \epsilon \to \mathcal{N}(0, \Sigma),$$

donde f define un proceso Gaussiano, podemos modelizar la distribución a priori como una normal, ya que $f(\mathbf{X})$ sigue una distribución normal multivariante. La distribución del modelo de observación mediante una regresión logística,

$$p(\mathbf{w}|f) = \left(\frac{1}{1+e^{-f}}\right)^{(1+\mathbf{w})/2} \left(\frac{1}{1+e^{f}}\right)^{(1-\mathbf{w})/2} \quad \forall \mathbf{w} = \{-1, 1\}.$$

Con estas dos distribuciones y la regla de Bayes podemos obtener una distribución a posteriori sobre los parámetros de f. Finalmente para una nueva predicción integraremos en todo el espacio de distribución de los parámetros.

Software utilizado para la realización de la práctica.

Para la realización de la practica he usado el sistema de cómputo numérico MatLab. Las funciones adicionales para el cálculo de procesos Gaussianos y la estimación de sus parámetros mediante distintos núcleos, se obtuvieron de GPML, `http://www.gaussianprocess.org/gpml/code/matlab/doc/`. Los núcleos usados fueron el de base radial,

$$\mathbf{K}(x,z) = (sf)^2 \exp\left(-\frac{(x-z)^T P^{-1}(x-z)}{2}\right),$$

donde P es la matriz identidad por el parámetro ell^2. Por lo tanto los parámetros de esta función núcleo son $hyp = (\log(ell), \log(sf))^T$. El lineal vendría definido por,

$$\mathbf{K}(x,z) = x \cdot z,$$

es decir el producto escalar de dos vectores, en este caso la función no tiene parámetros adicionales.

Para obtener los parámetros óptimos del núcleo de base radial hemos usado la función minimize() de GPML. La función minimize() minimiza una función diferenciable usando el método de los gradientes conjugados.

Resultados experimentales.

El problema de clasificación está altamente desbalanceado, el número de instancias que presentan ser cancerígenas es aproximadamente cuatro veces menor de las instancias sanas. Para lidiar el problema usaremos una técnica de bagging, para cada fold construiremos cuatro subconjuntos disjuntos de entrenamiento y uno solo de test. Estos cuatro subconjuntos serán de aproximadamente el mismo tamaño y deben cubrir todo el conjunto de instancias en el fold.

```
%Creamos cada uno de los Folds con sus respectivos conjuntos train y test.
%Fold1: 1Test y 4Train.
Fold1_Test = [Malign_folds(1).histogram; Healthy_folds(1).histogram];

Fold1_Train_1 = [Malign_folds(2).histogram; Healthy_folds(2).histogram];

Fold1_Train_2 = [Malign_folds(3).histogram; Healthy_folds(3).histogram];

Fold1_Train_3 = [Malign_folds(4).histogram; Healthy_folds(4).histogram];

Fold1_Train_4 = [Malign_folds(5).histogram; Healthy_folds(5).histogram];
```

Figura 2: Construcción de conjuntos train y test en un mismo fold.

Repetimos el proceso para cada uno de los folds. Obtendremos así un total de 20 clasificadores, cuatro por cada fold. Los clasificadores nos dan la probabilidad de que una imagen presente cáncer, calcularemos la probabilidad media de cáncer en cada fold.

Computo curvas ROC.

Con estas probabilidades medias calcularemos las curvas ROC y Precisión-Recall. Lo haremos mediante la función perfcruve(). Para el fold1 obtendríamos las siguientes gráficas, los próximos cálculos están hechos para un proceso Gaussiano de base radial:

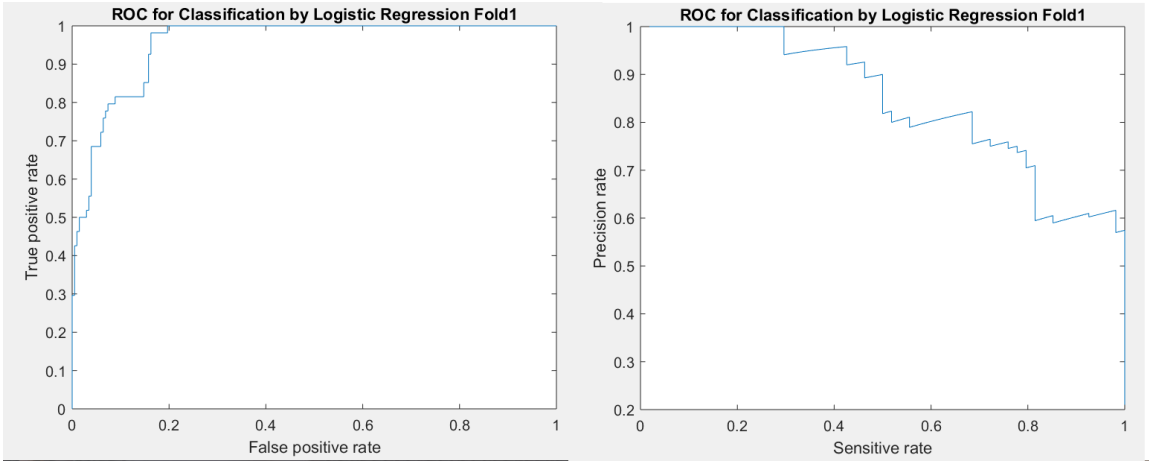

Figura 3: A la izquierda curva ROC, derecha Precisión-Recall.

El plano dado para la curva ROC viene definido la especifidad, eje abscisas, y la sensibilidad, eje ordenadas. Cada salto en la el gráfico representa un valor distinto en el umbral de clasificación, por tanto la visualización de este tipo de curvas nos ayuda a elegir un θ más adecuado al tipo de clasificación que queramos. En este caso que se trata de decidir entre presencia o ausencia de cáncer, queremos reducir al máximo el número de falsos negativos.

La curva Precisión-Recall funciona de la misma forma que la ROC pero en este caso el eje de abscisas viene representado por la sensibilidad y el de ordenadas por la precisión. A continuación mostramos las curvas obtenidas para los folds: 2, 3, 4 y 5 respectivamente.

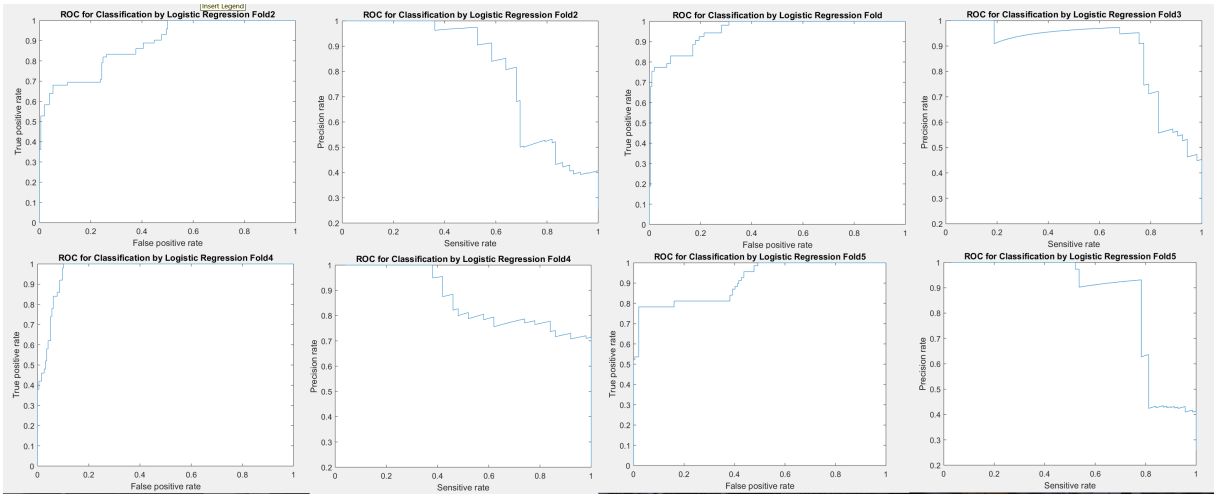

Figura 4: Curvas ROC y Precision-Recall para cada uno de los folds.

Una medida de bondad del clasificador sería el área debajo de estas curvas, un valor muy cercano a 1 sería lo ideal. Los valores obtenidos son los siguientes:

	Fold1	Fold2	Fold3	Fold4	Fold5
Área ROC	0.9528	0.8825	0.9549	0.9670	0.9102
Área PR	0.8259	0.7896	0.8566	0.8267	0.8316

Cuadro 1: Áreas bajo las curvas ROC y Precision-Recall, kernel radial.

De forma similar calculamos las probabilidades medias dadas por un proceso Gaussiano con núcleo lineal, obteniendo así los siguientes resultados para las áreas de las curvas ROC y Precision-Recall.

	Fold1	Fold2	Fold3	Fold4	Fold5
Área ROC	0.8134	0.7099	0.9489	0.8841	0.9267
Área PR	0.6491	0.5103	0.8362	0.8248	0.8223

Cuadro 2: Áreas bajo las curvas ROC y Precision-Recall, kernel lineal.

Computo métricas de bondad.

Tomaremos como umbral $\theta = 0{,}5$ para la clasificación. Calcularemos la matriz de confusión y con ello las distintas mátricas de bondad para nuestro clasificador. La matriz de confusión del modelo obtenido con un núcleo radial para el fold1 viene dada por:

		Predicciones	
		Cáncer	No Cáncer
Etiquetas	Cáncer	30(VP)	24(FN)
	No Cáncer	7(FP)	196(VN)

A partir de esta tabla podemos calcular las siguientes medidas para el fold1:

- Accuracy: $\frac{VP+VN}{VP+FN+FP+VN} = 0{,}8794$.
- Specifity: $\frac{VN}{VN+FP} = 0{,}9655$.
- Recall: $\frac{VP}{VP+FN} = 0{,}5556$.
- Preccision: $\frac{VP}{VP+FP} = 0{,}8108$.
- F1 score: $\frac{2VP}{2VP+FP+FN} = 0{,}6593$.

En la siguiente tabla se recogen todos los datos obtenidos para uno de los folds.

	Fold1	Fold2	Fold3	Fold4	Fold5
Accuracy	0.8794	0.7447	0.9305	0.8821	0.9104
Specifity	0.9655	0.7619	0.9951	0.9592	0.9548
Recall	0.5556	0.6944	0.6792	0.5800	0.7826
Preccison	0.8108	0.5000	0.9730	0.7838	0.8571
F1	0.6593	0.5814	0.8000	0.6667	0.81571

Cuadro 3: Métricas de bondad en los distintos folds, kernel radial.

Realizamos el mismo estudio pero con un núcleo lineal, obteniendo así:

	Fold1	Fold2	Fold3	Fold4	Fold5
Accuracy	0.7899	0.7447	0.7954	0.7967	0.7425
Specifity	1	1	1	1	1
Recall	0	0	0	0	0
Preccison	Nan	Nan	Nan	Nan	Nan
F1	Nan	Nan	Nan	Nan	Nan

Cuadro 4: Métricas de bondad en los distintos folds, kernel lineal.

Discusión de los resultados.

Comenzaremos con los resultados obtenidos para el kernel lineal, por lo general el área de la curva ROC es alto, lo cual es una buena señal del clasificador. Más de lo mismo para el área Precision-Recall aunque esta en los folds 1 y 2 no obtiene unos muy buenos resultados. Con estas medidas no sabemos mucho acerca de como de bueno es el clasificador, los valores del área de las curvas no son suficientemente significativos para dar una conclusión.

La cosa cambia cuando vemos que ocurre con las métricas que hemos calculado de la matriz de confusión. Una specifity igual a 1 nos indica que no hemos cometido ningún falso postivo, por el contrario un reacall igual a 0 informa que no hemos predicho ninguna imagen cancerígena. Por lo tanto podemos concluir que todos los valores que devuelve el clasificador es ausencia de cáncer en todas las imágenes. Se desecha por tanto el uso de núcleo lineal para este problema.

Ahora analizaremos los datos recogidos para el núcleo radial. Para todos los folds el área debajo de las curvas es alta. Fijándonos en las gráficas para las curvas ROC estas se ven muy próximas a loa que rsería la curva ideal, para la Precision-Recall se observan saltos más grandes respecto un cambio en el umbral de clasificación. Con ello podemos decidir si un cambio del umbral nos dará una clasificación más acorde a nuestro problema, aquí por ejemplo queremos reducir al máximo el número de falsos negativos; es decir, predecir que asuencia de cáncer cuando realmente hay presencia.

En los datos recogidos para un umbral $\theta = 0{,}5$ observamos que los valores más bajos se obtienen para el Recall, lo cual índica una presencia alta de falsos negativos. Specifity que tiene valores muy altos muestra la poca presencia de falsos positivos. Los valores de precision son bastante buenos, la predicción de Cáncer es positiva. F1 es una media armónica entre la precision y specifity.

Clasificación para nuevas imágenes.

Para una nueva imagen le calcularemos bloques de tamaño 2048*2048, a cada uno de estos bloques le calcularemos un descriptor lbp uniforme como se hizo en los datos proporcionados. Ahora necesitamos usar todo nuestro conjunto de datos para entrenar el clasificador, por lo que no desecharemos un 20 % de los datos para dedicarlos al test.

Para todos los datos crearemos cuatro subconjuntos disjuntos, cada uno con un número similar de imágenes con presencia de cáncer y ausencia. Entrenaremos un clasificador con procesos gaussianos para cada uno de estos conjuntos disjuntos, lo primero que haremos será ver las curvas ROC y Precison-Recall para decidir el mejor umbral de clasificación, este caso el que reduce el número de falsos negativos.

Una vez decidido el umbral de decisión, tomaremos la probabilidad media que nos devuelven los cuatro clasificadores y con ello calcularemos la etiqueta de ese bloque.

Finalmente habremos obtenido distintos bloques en la imagen con presencia o ausencia de cáncer, por lo que podremos detectar en que zonas de imagen hay presencia de cáncer.

Diseño de experimento adicional.

En la práctica descrita hemos utilizado bagging para balancear las clases. Tenemos 1014 y 298 ejemplos sanos y cancerígenos, respectivamente. Podemos pensar en otras técnicas que ayuden al clasificador con el problema del desbalance de clases.

Nos gustaría crear nuevas instancias de la clase minoritaria, para ello la pregunta sería, ¿De que forma podemos crear estas nuevas instancias? Una primera idea intuitiva, si el descriptor de una imagen presenta cáncer, probablemente un descriptor muy parecido debería representar una imagen con presencia de cáncer. Para hacer esto nos fijaremos en los k vecinos más cercanos de una instancia con presencia de cáncer, de esos k tomamos j de forma aleatoria y y crearemos una nueva instancia haciendo una interpolación entre esos datos.

Una vez los datos estén balanceados, dividiremos nuestro conjunto de datos en 5 folds, haremos esta división de forma estratificada, para que cada uno de estos tenga las mismas instancias y la misma carga de desbalance en las clases, además estos folds son muestras representativas del conjunto total de datos. Con ello haremos una validación cruzada sobre el clasificador que vamos a entrenar y testear.

Haremos este proceso con distintos parámetros del clasificador, cambiaremos los núcleos, parámetros de los núcleos o incluso usar otro tipo de clasificador como svm o árbol de decisión. Una vez tengamos los valores para cada fold con los distintos modelos, compararemos los resultados con test estadísticos. Como hemos usado fuertemente la hipótesis de normalidad para la construcción de los procesos gaussianos, seguirmeos usando la hipótesis para hacer test paramétricos. Mediante un test ANOVA podremos determinar que modelo es el mejor para la clasificación de imágenes histológicas.

Bibliografía

[1] Simon D. J. Prince, *Computer Vision: Models, Learning and inference.*, http://www.computervisionmodels.com/

[2] Wikipedia, *Bayesian Inference.* https://en.wikipedia.org/wiki/Bayesian_inference

[3] MathWorks, *Centro de Ayuda, perfcurve().* https://es.mathworks.com/help/stats/perfcurve.html