TP3 - Reconocimiento de Patrones

Elio Campitelli

SVM con separación lineal

1. Implementar el algoritmo Pegausus y probarlo en datos sintéticos en R^2 con dos clases utilizando soft margin y separación lineal. Esta clases pueden ser generadas con distribución Gaussiana bivariada con covarianza isotrópica.

2. Utilizando el algoritmo y los datos del punto anterior calcular el error de clasificación para distintos valores de C y graficar, para cada caso, la recta de separación del clasificador ($\mathbf{w}^t\mathbf{x} + x_0 = 0$), así como también las rectas correspondientes a $\mathbf{w}^t\mathbf{x} + x_0 = 1$ y a $\mathbf{w}^t\mathbf{x} + x_0 = -1$. Hacer esto para 3 diferentes grados de separación entre las dos clases.

Voy a implementar SVM usando soft-margin para datos sintéticos generados a partir de dos distribuciones normales bivariadas con $\Sigma^2 = I$ y distintos grados de separación ($\delta \in 1, 2, 4, 6$) en dirección x. La recta de separación óptima es una recta vertical centrada en cero.

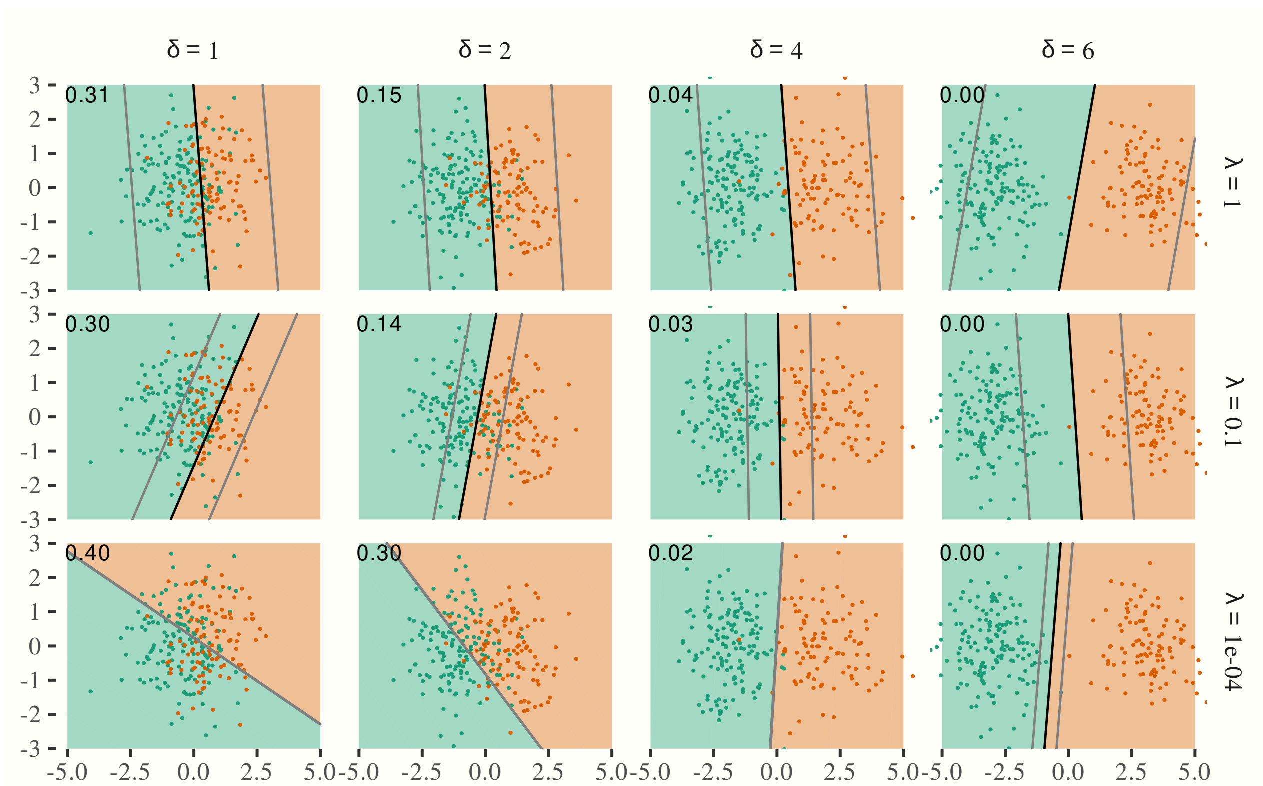

Figure 1: Límites de decisión para datos sintéticos con distintos grados de separación (columnas) y distinto nivel de tolerancia para datos mal clasificados (filas). La línea negra es la recta de separación y las líneas grises son el márgen $\mathbf{w}^t\mathbf{x} + x_0 = \pm 1$. El error de clasificación (proporción de datos mal clasificados) se muestra an la esquina superior izquierda.

En la Figura 1 se muestra el límite de decisión para los cuatro set se datos (columnas) y los grados de soft-margin donde un λ pequeño equivale a poca tolerancia mientras que un lambda grande equivale a mucha tolerancia.

SVM con kernel gaussiano

3. Disenãr un generador de datos no separables y utilizar el kernel $K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma||\mathbf{x} - \mathbf{x}'||)$ para separar un set de datos producido por este generador.

Luego, extiendo el algoritmo PEGASOS para poder usar un kernel arbitrario. En la Figura 2 se muestran los resultados de applicar SVM con kernel gaussiano con distintos valores de γ y λ y un $\delta = 3.5$.

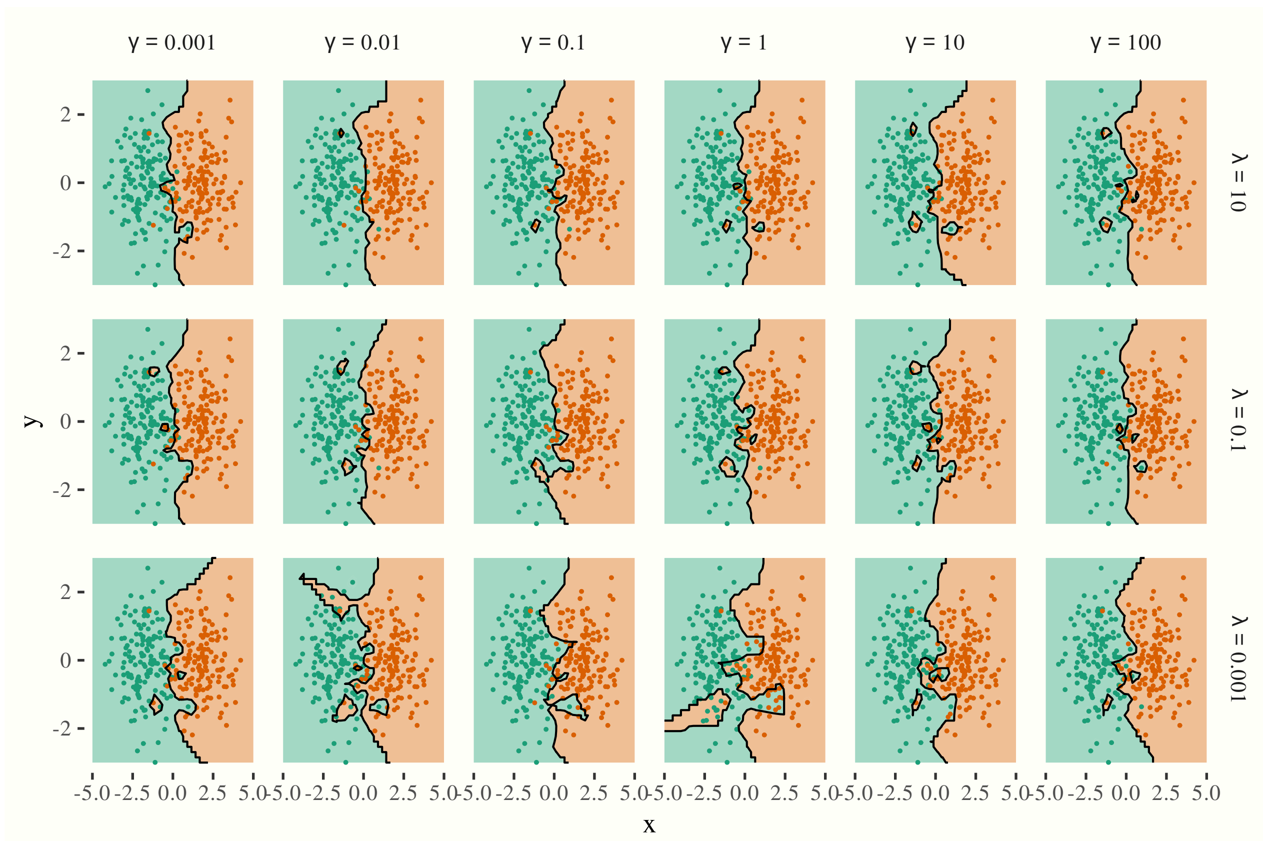

Figure 2: SVM con kernel gaussiano para distintos valores de γ y λ. La separación entre los datos es $\delta = 2$.

Estos datos sintéticos son un poco aburridos. Voy a usar los datos que se muestran en la Figura 7.2 del Bishop (Figura 3 derecha). Digitalicé los datos y los usé para entrenar una SVM con kernel gausiano con $\gamma = 0.5$ y $\lambda = 0.1$. El resultado se muestra en la Figura 3 izquierda, que es muy similar.

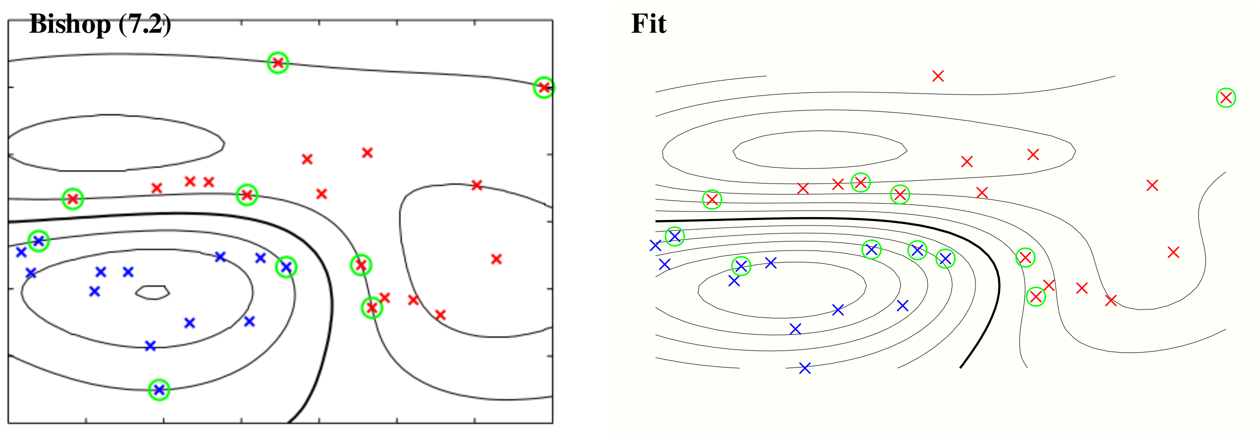

Figure 3: SVM con kernel gaussiano para los datos de la Figura 7.2 del Bishop. y(x) en líneas de contorno, y(x) = 0 en línea gruesa. A la derecha, la figura original, a la izquierda, el clasificador entrenado en este ejercicio

Validación cruzada

4. Utilizando el mismo generador de datos del item anterior, determinar la mejor combinación de valores (C, γ) mediante cross-validation.

El panel izqueirdo de la Figura 4 muestra la tasa de acierto (proporción de observaciones correctamente categorizadas) computada por validación cruzada para distintos valores de λ y γ. La mejor combinación de hiperparámetros está marcada con un punto. En el panel derecho se muestra el límite de decisión para la SVM ajustada con éstos.

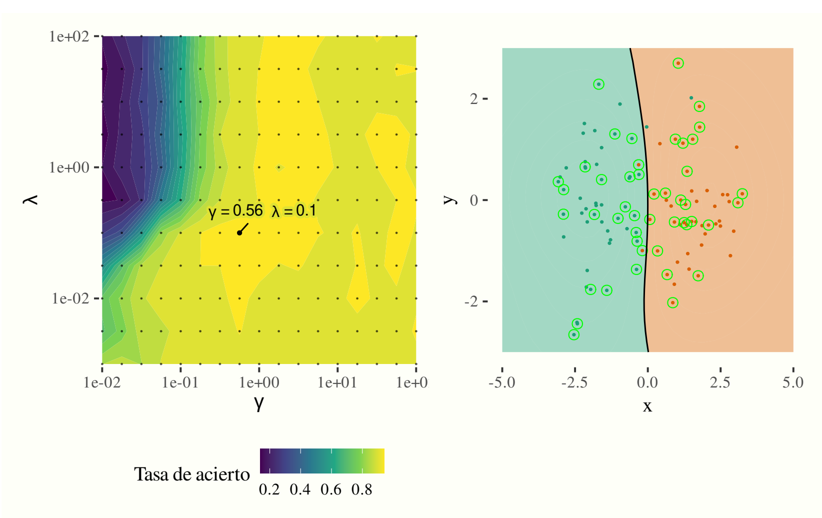

Figure 4: Acierto computado a partir de validación cruzada para distintos valores de λ y γ y SVM fiteada con los mejores valores de λ y γ elegidos por validación cruzada.