Clase 7: Evaluación de clasificadores y generalizaciones de la regresión

Simón Leiva Meza

Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Programa de Magister en Estadística

Gráfico de Ganancia

Es de lo gráficos más interpretables, indica cuanto ganamos al usar nuestro modelo para gestionar al $x\%$ del tarjet.

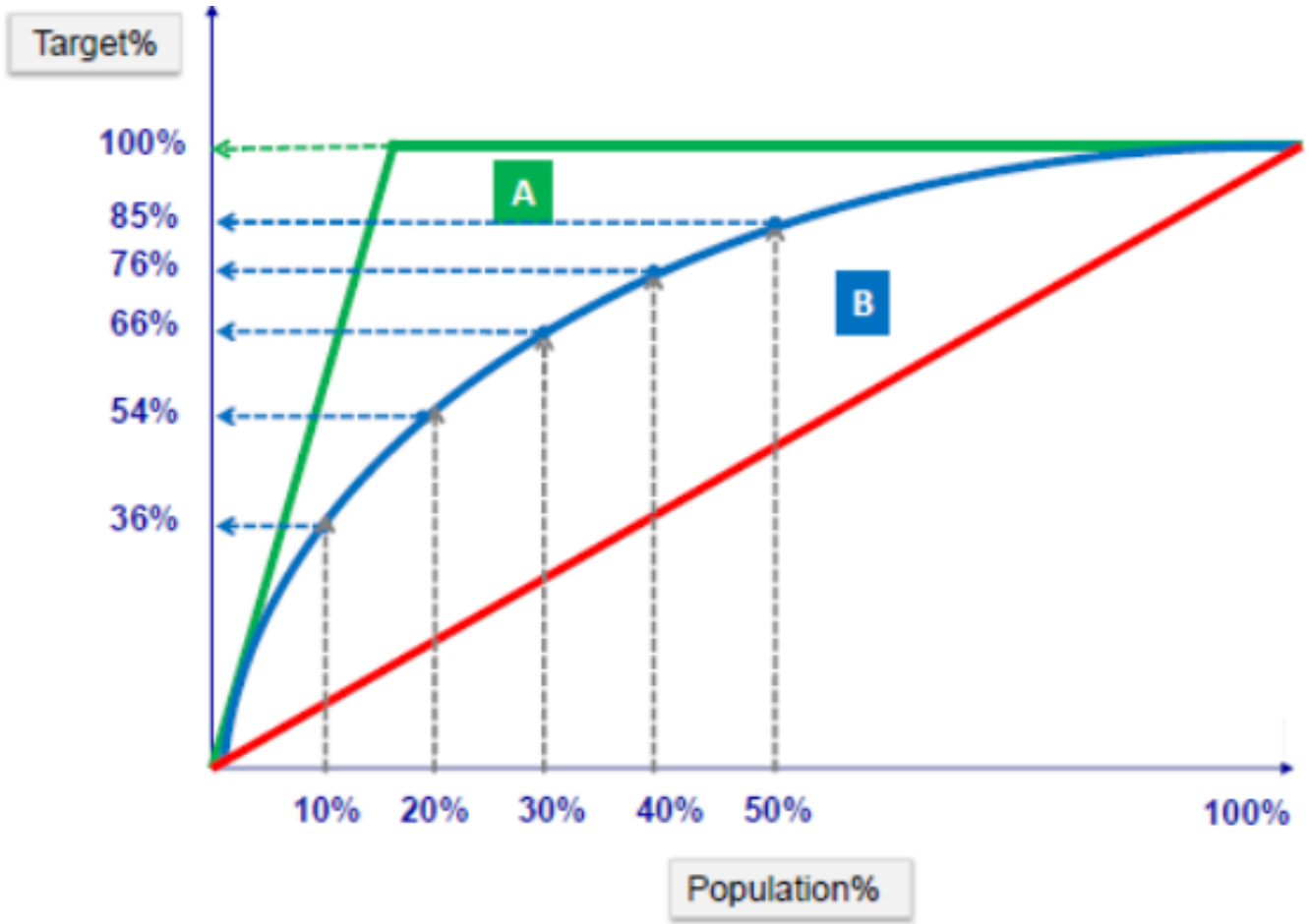


Gráfico de Ganancia

A diferencia de la ROC, este gráfico no se construye a partir de la matriz de confusión, sino más bien a partir de las mismas predicciones

Construcción de gráfico de Ganancia

1. En la data de test, ordenar los valores reales y predichos de forma descendente según las probabilidades estimadas.
2. Graficar en el eje X el orden de los datos escalado entre 0 y 100.
3. Graficar en el eje Y el acumulado de unos reales.
4. Para la curva verde, graficar el caso del modelo perfecto, esto es ordenar según valores reales y repetir el proceso.

El KPI que normalmente resume este gráfico es que porcentaje de los casos toma el 10% más propenso.

Gráfico de Ganancia

Resultado	Probabilidad
1	0,67
1	0,43
1	0,89
1	0,78
1	0,52
1	0,32
1	0,73
1	0,41
1	0,92
1	0,21
0	0,12
0	0,43
0	0,32
0	0,69
0	0,83
0	0,33
0	0,41
0	0,27
0	0,15
0	0,52

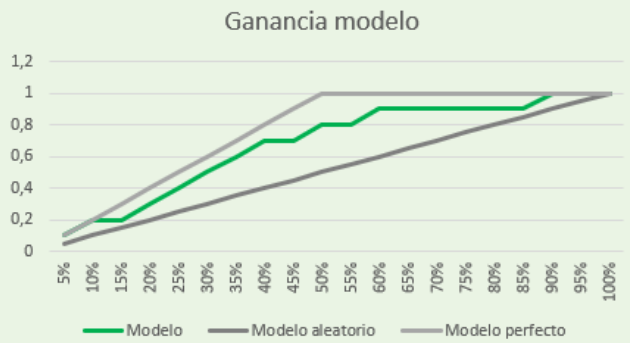
1

Se comienza con un resultado que ya conocemos (por ejemplo el test) y las probabilidades predichas por nuestro modelo. Sólo necesitamos eso para calcular el Lift.

Resultado	Probabilidad	Modelo	Modelo aleatorio	Modelo perfecto
1	0,92	0,1	0,05	0,1
1	0,89	0,2	0,1	0,2
0	0,83	0,2	0,15	0,3
1	0,78	0,3	0,2	0,4
1	0,73	0,4	0,25	0,5
1	0,67	0,5	0,3	0,6
1	0,61	0,6	0,35	0,7
1	0,52	0,7	0,4	0,8
0	0,52	0,7	0,45	0,9
1	0,43	0,8	0,5	1
0	0,43	0,8	0,55	1
1	0,41	0,9	0,6	1
0	0,41	0,9	0,65	1
0	0,37	0,9	0,7	1
0	0,33	0,9	0,75	1
0	0,32	0,9	0,8	1
0	0,27	0,9	0,85	1
1	0,21	1	0,9	1
0	0,15	1	0,95	1
0	0,12	1	1	1

2

Se ordenan las probabilidades estimadas de forma descendente y se obtiene el porcentaje acumulado de los unos.

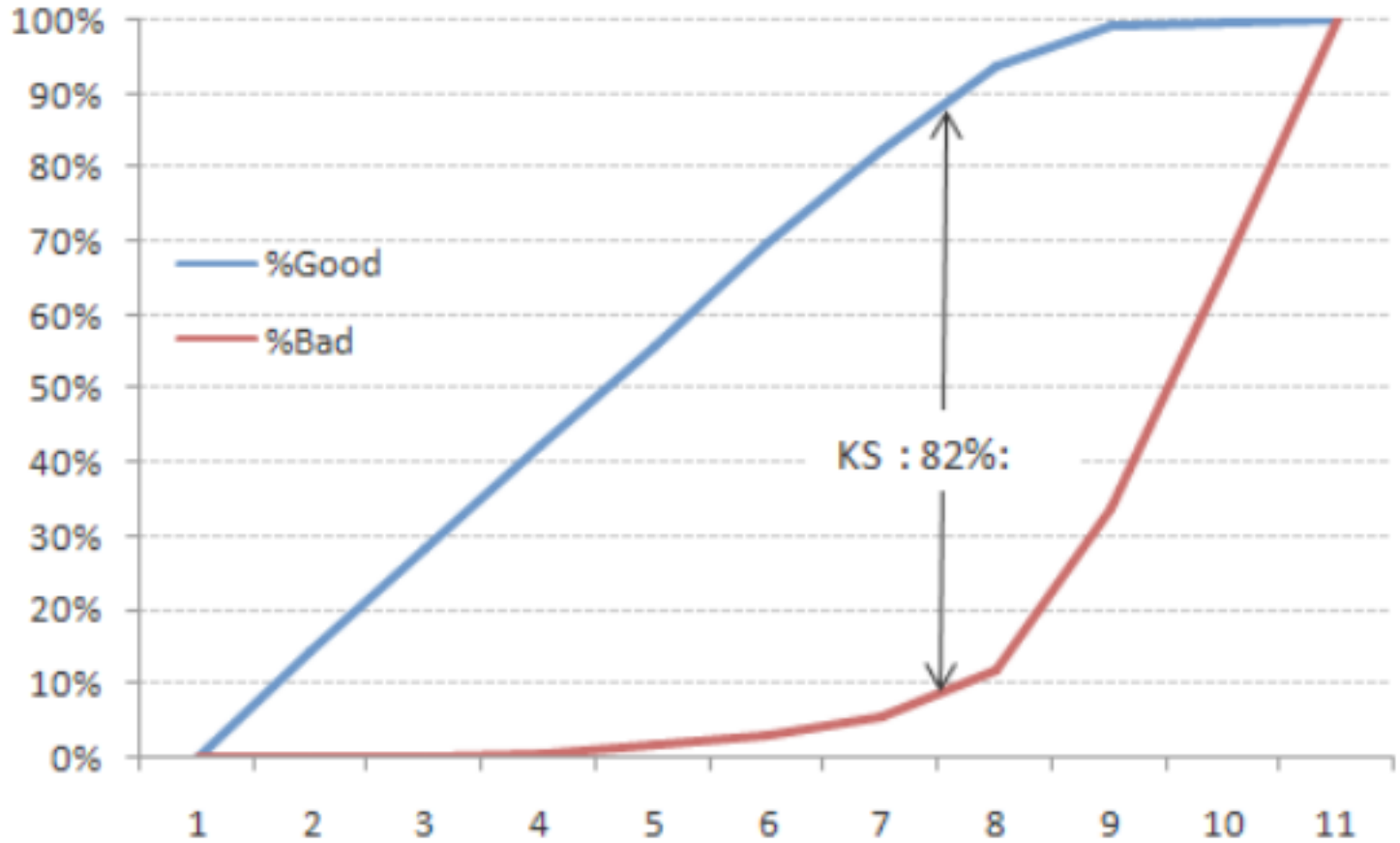


3

Se grafican el porcentaje acumulado (línea verde) una recta (línea ploma) y como sería el modelo capaz de captar óptimamente el orden de probabilidad (línea gris). En general se calcula que porcentaje de ganancia se obtiene gestionando el 10% o 20% con probabilidad más alta. En este ejemplo, tomando el 10% según el modelo, logramos captar el 20% de los que de verdad son unos.

Gráfico KS

El gráfico KS (Kolmogorov-Smirnov) se basa en el famoso test. Lo que busca es comparar que tan parecidas son las distribuciones acumuladas de los casos (1) versus los no casos (0)



Construcción Gráfico KS

La construcción es similar al gráfico de ganancia.

Construcción de gráfico KS

1. En la data de test, ordenar los valores reales y predichos de forma descendente según las probabilidades estimadas.
2. Graficar en el eje Y el acumulado de unos reales.
3. Graficar en el eje Y el acumulado de los ceros reales
4. Graficar en el eje X el orden de los datos escalado entre 0 y 100.

El KPI resumen tiene su mismo nombre, es la distancia máxima entre las dos curvas. Un KS bajo es señal de que no capacidad de discriminación entre ambas clases y uno alto es que hay mucha separación,

Resultado	Probabilidad
1	0,67
1	0,43
1	0,89
1	0,78
1	0,52
1	0,32
1	0,73
1	0,41
1	0,92
1	0,21
0	0,12
0	0,43
0	0,32
0	0,69
0	0,83
0	0,33
0	0,41
0	0,27
0	0,15
0	0,52

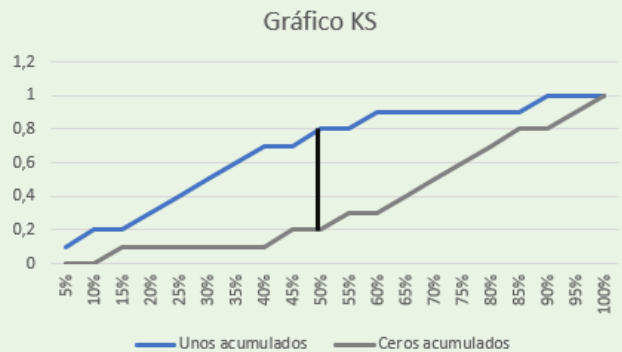
1

Se comienza con un resultado que ya conocemos (por ejemplo el test) y las probabilidades predichas por nuestro modelo. Sólo necesitamos eso para calcular el KS.

Resultado	Probabilidad	Unos acumulados	Ceros acumulados
1	0,92	0,1	0
1	0,89	0,2	0
0	0,83	0,2	0,1
1	0,78	0,3	0,1
1	0,73	0,4	0,1
1	0,67	0,5	0,1
1	0,61	0,6	0,1
1	0,52	0,7	0,1
0	0,52	0,7	0,2
1	0,43	0,8	0,2
0	0,43	0,8	0,3
1	0,41	0,9	0,3
0	0,41	0,9	0,4
0	0,37	0,9	0,5
0	0,33	0,9	0,6
0	0,32	0,9	0,7
0	0,27	0,9	0,8
1	0,21	1	0,8
0	0,15	1	0,9
0	0,12	1	1

2

Al igual que en el gráfico de ganancia, se ordenan de la misma forma pero además del porcentaje de unos acumulados, se calcula el porcentaje de ceros acumulados.

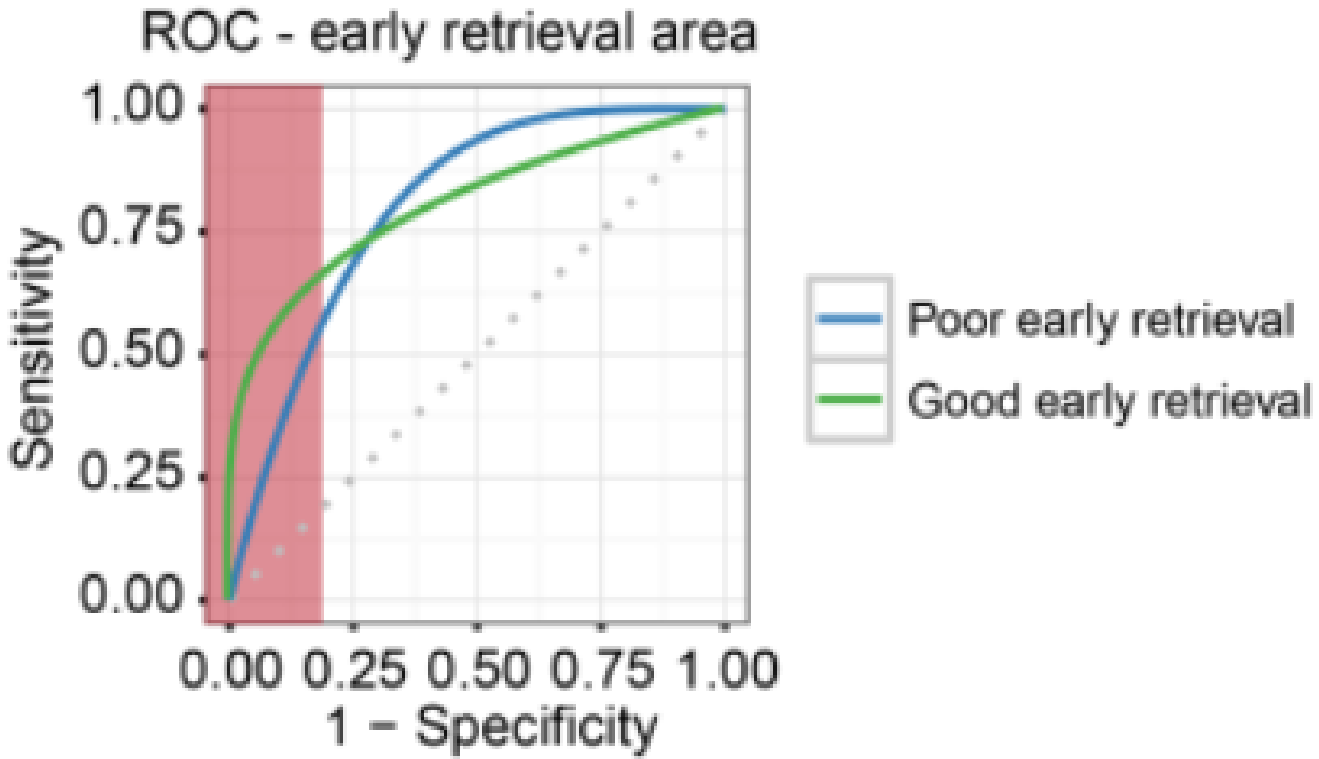


3

Se grafican ambas curvas, el de los unos acumulados y de los ceros acumulados. Mientras más separadas estén las curvas mejor. La medida resumen es el indicador KS que es la distancia máxima entre las dos curvas, en este ejemplo el KS es de 0,6

Curva CROC

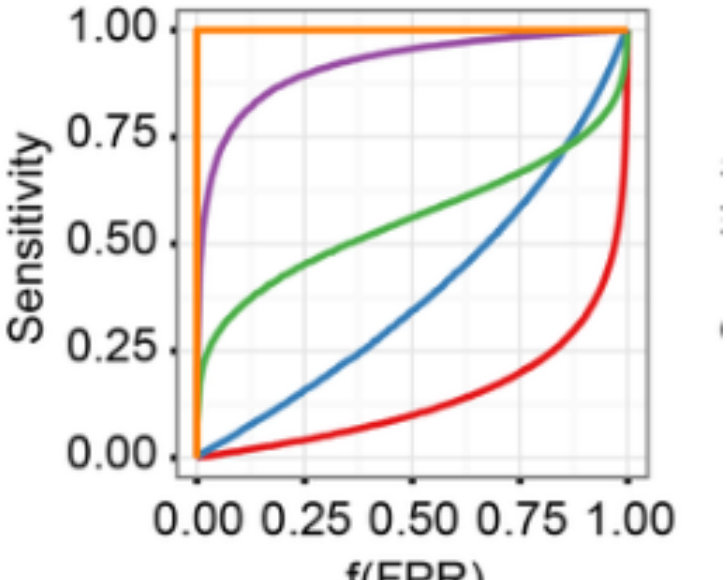
La curva de ROC de los indicadores más populares pero puede fallar su interpretación, pues no es tan directa como el lift o el KS. En particular dos curvas con igual AUC nos gustaría la que tenga mejor forma. Esta es la que a baja tasa de falso positivo tiene alta tasa de verdadero positivo. Esto se llama retorno temprano.



Curva CROC

Para solucionar este problema se propuso una alteración sencilla a la curva, esto se conoce como Concentrated Receiver Operating Characteristic (CROC) (Swamidass, 2010). Consiste en simplemente aplicar una función concentradora a la tasa de falsos positivos, antes de la construcción de la curva, de esta forma se exagera a la vista el retorno temprano de la curva

$$f(FPR) = \frac{1 - \exp(-\alpha \cdot FPR)}{1 - \exp(-\alpha)}$$



Curva PR

Al igual que su hermano, la curva de ROC, la curva PR (Precision-Recall), está basada en la matriz de confusión, pero en lugar de graficar la tasa de falso positivo contra la de verdadero positivo, grafica el trade off entre la precisión y el recall. Recordar que un buen clasificador tendrá alta la precision y recall

1. Si hay muchos más ceros que unos, es probable obtener muchos falsos positivos, por tanto una precisión baja, independiente de la cantidad de verdaderos positivos.
2. Si hay muchos más unos que ceros, es probable obtener muchos falsos negativos, por tanto un recall bajo, independiente de la cantidad de verdaderos positivos.

La curva de ROC sólo grafíca el Recall (o sensibilidad) y el fall-out (1-especificidad o tasa de falso positivo), el cual tiende a aligerar el efecto de un desborde en los falsos positivos, en consecuencia la curva de ROC tiende a ser insensible a desbalances.

Curva PR

La curva PR se construye de la siguiente forma

Construcción de la curva PR

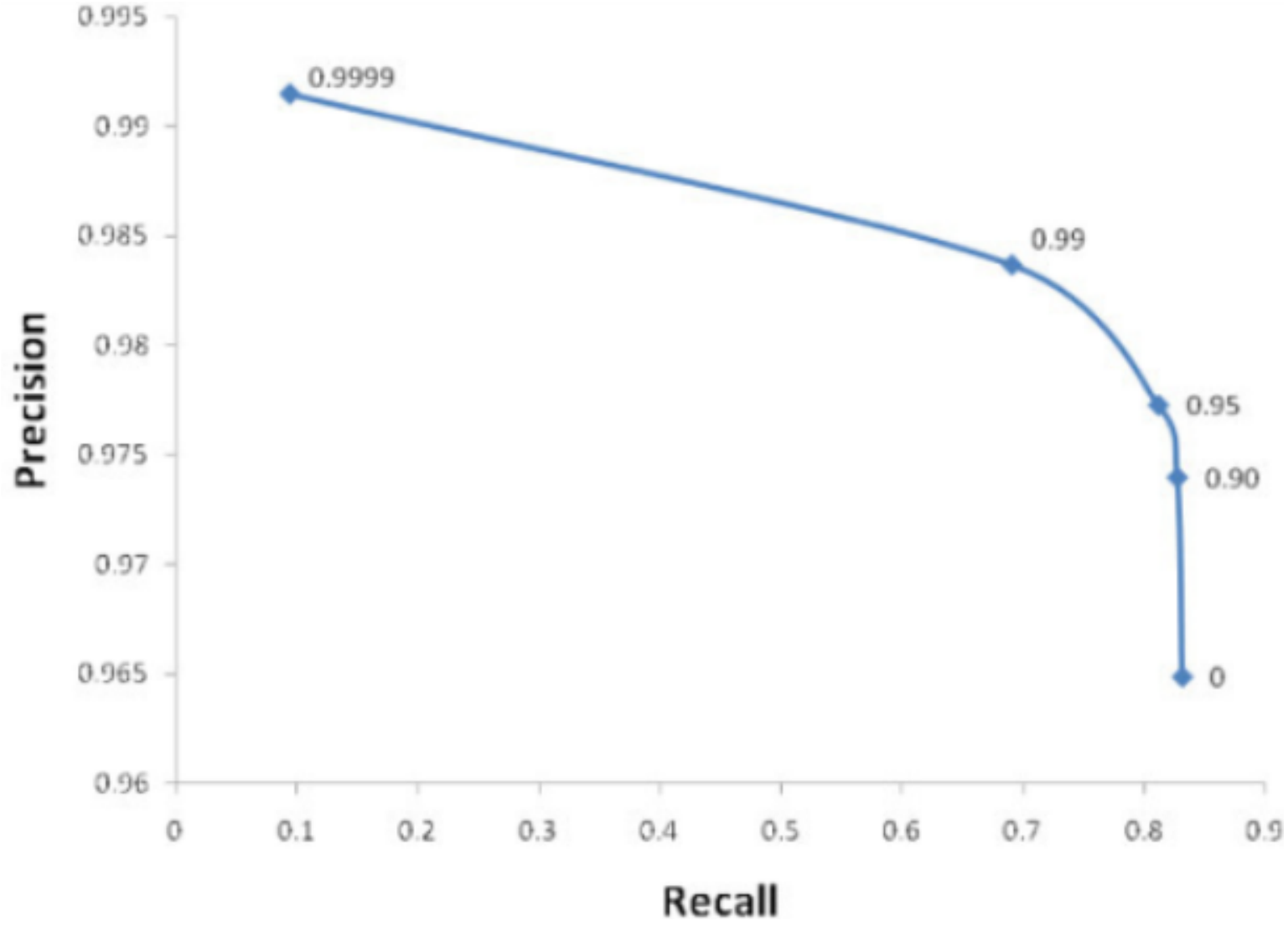
1. En la data de test, obtener la precisión y el recall para n cortes de probabilidad.
2. Graficar en el eje Y las precisiones obtenidas.
3. Graficar en el eje X los recall obtenidos.

El KPI asociado a este gráfico es el AUC, calculado como:

$$AUC = \sum_{i=2}^{n} (R_i - R_{i-1}) \cdot P_i$$

A diferencia de la ROC, este se mueve entre cero y uno.

Curva PR


0.995
0.99
0.985
0.98
0.975
0.97
0.965
0.96
Precision
0.9999
0.99
0.95
0.90
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Recall

Ejemplo

Con los datos de loan, ajuste una regresión logística sencilla (como el de la clase pasada). Ajuste con datos sin balancear. Obtenga la ROC y la curva PR. Compare.

Comentarios

1. El gráfico de ganancia y el KS es preferible usarlo en situaciones que es necesario interpretar resultados de cara a gente no especializada.
2. Normalmente el ROC es de los mejores indicadores, pero es necesario interpretar la curva entera, el CROC ayuda a facilitar la interpretación.
3. El gráfico PR es recomendable su uso en casos de desbalance o bien que la muestra sea vulnerable a muchos falsos positivos o falsos negativos.

Modelos Aditivos Generalizados

Repaso

1. Modelo Lineal (LM)

$$\mu_i = X_i\beta, \quad \mu_i = E(Y_i), \quad Y_i \sim N(\mu_i, \sigma^2)$$

2. Modelo Lineal Generalizado (GLM)

$$g(\mu_i) = X_i\beta, \quad \mu_i = E(Y_i), \quad Y_i \sim EF(\mu_i, \phi)$$

3. Modelo Aditivo Generalizado (GAM)

$$g(\mu_i) = A_i\beta + \sum_j f_j(x_{ji}), \quad \mu_i = E(Y_i), \quad Y_i \sim EF(\mu_i, \phi)$$

Modelos Aditivos Generalizados

En la formula anterior, f_j son funciones de suavizamiento lineales. En términos simples un GAM es un GLM tomando como predictores un conjunto de combinaciones lineales de funciones.

Bajo esta nueva especificación, se deben resolver dos problemas fundamentales: la elección de las funciones de suavizamientos y elegir qué tan suaves debieran ser. En general las funciones de suavizamiento que se tienden a elegir son combinaciones lineales de polinomios, sin embargo estos mismos tienen el mismo problema que una serie de Taylor, sólo predicen bien en una vecindad.

Modelos Aditivos Generalizados

Polinomios a tramos

Definamos un primer modelo de con la siguiente base:

$$h_1(X) = I(X < \xi_1), \quad h_2(X) = I(\xi_1 \leq X < \xi_2), \quad h_3(X) = I(\xi_2 < X)$$

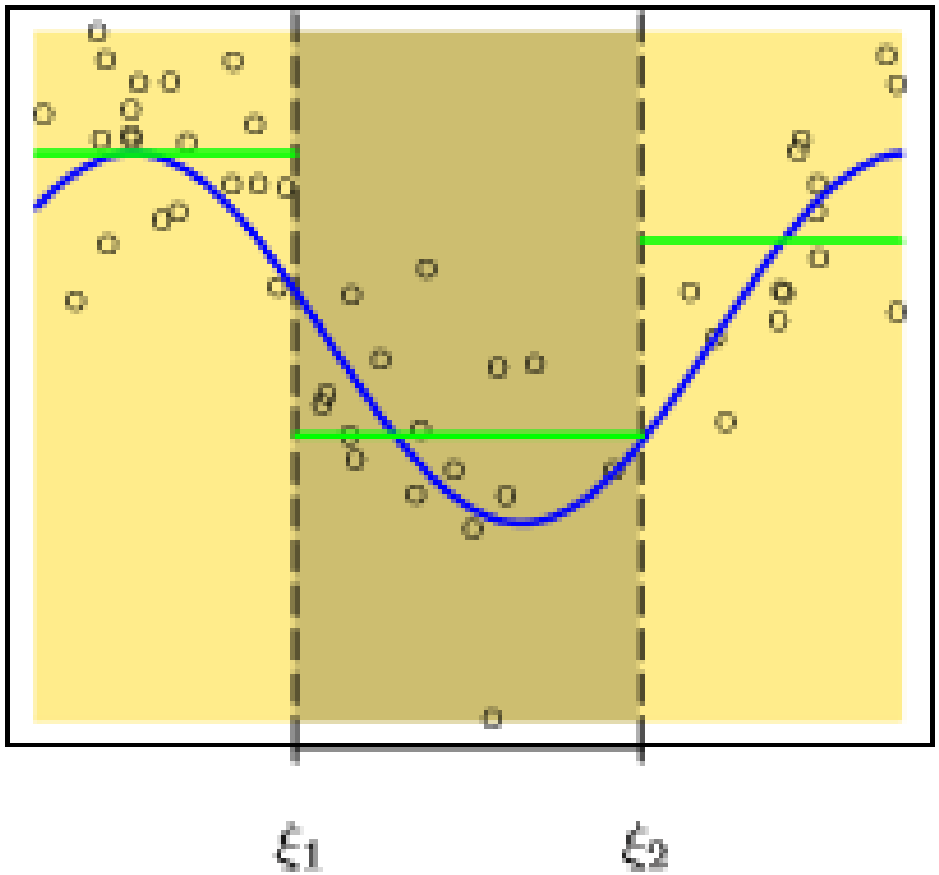
Luego, definimos la función suavizadora como combinación lineal de estas bases

$$f(X) = \sum_{m=1}^{3} \beta_m h_m(X)$$

Los valores ξ_j desde ahora en adelante les llamaremos nodos.

Modelos Aditivos Generalizados

Polinomios a tramos



Modelos Aditivos Generalizados

Polinomios a tramos

El siguiente grado de refinamiento es que las funciones por tramos tengan pendiente, lo cual es trivial de lograr, simplemente es cosa de agregar tres bases más

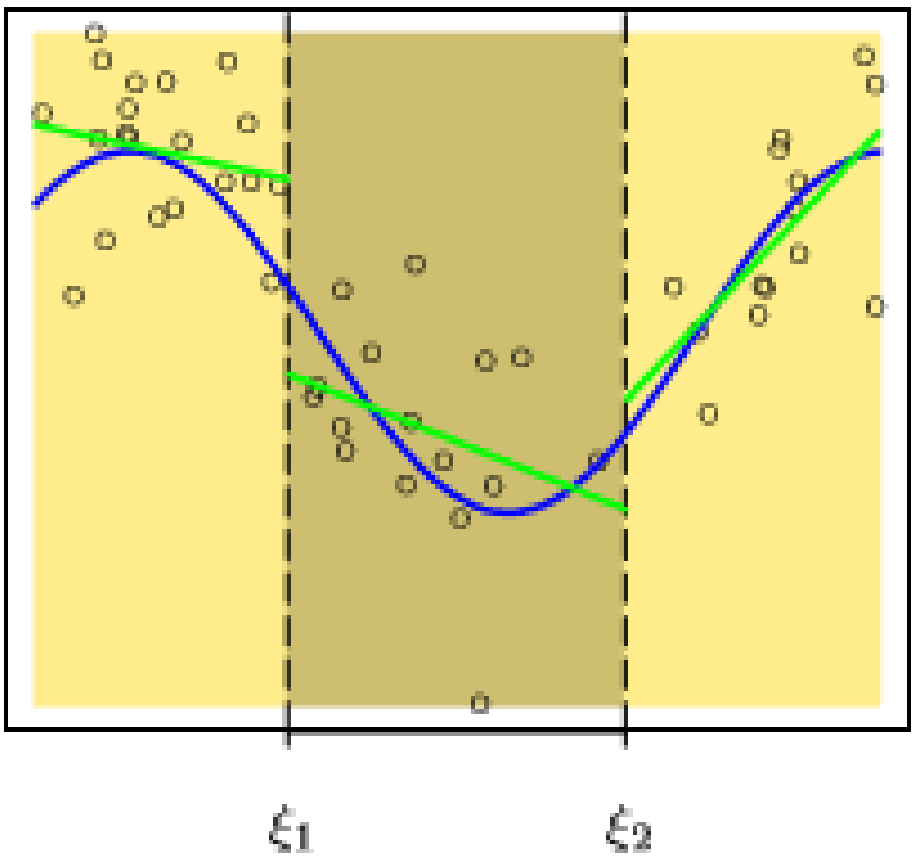
$$h_1(X) = I(X < \xi_1), \quad h_2(X) = I(\xi_1 \leq X < \xi_2), \quad h_3(X) = I(\xi_2 < X)$$

$$h_4(X) = I(X < \xi_1) \cdot X, \quad h_5(X) = I(\xi_1 \leq X < \xi_2) \cdot X,$$

$$h_6(X) = I(\xi_2 < X) \cdot X$$

Modelos Aditivos Generalizados

Polinomios a tramos



Modelos Aditivos Generalizados

Polinomios a tramos

Ciertamente es más preciso que el caso anterior, sin embargo una de las cualidades deseables que debiera tener un predictor es la continuidad, es decir $f(\xi_1^-) = f(\xi_1^+)$. Tomando las bases anteriores, esto implica que:

$$\beta_1 + \xi_1\beta_4 = \beta_2 + \xi_1\beta_5$$

y

$$\beta_2 + \xi_1\beta_5 = \beta_3 + \xi_1\beta_6$$

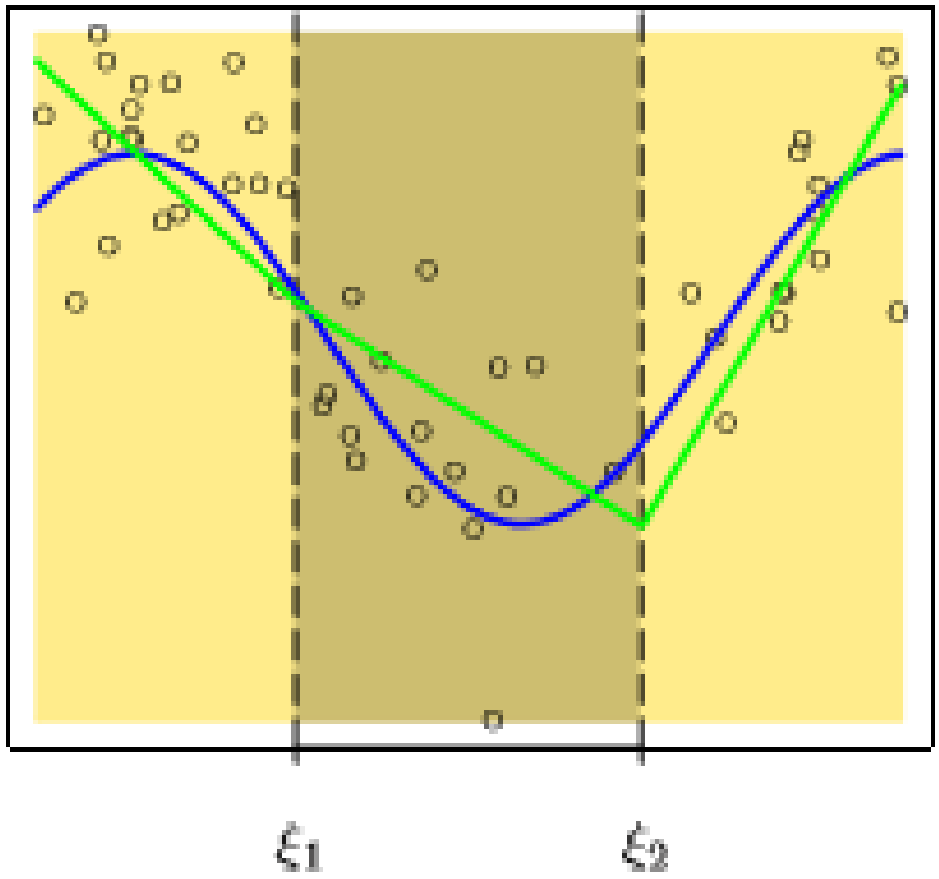
Luego, el caso continuo se puede representar de forma más elegante como:

$$h_1(X) = 1, \quad h_2(X) = X, \quad h_3(X) = (X - \xi_1)_+, \quad h_4(X) = (X - \xi_2)_+$$

donde h_+ representa la parte positiva de h

Modelos Aditivos Generalizados

Polinomios a tramos



Modelos Aditivos Generalizados

Polinomios a tramos

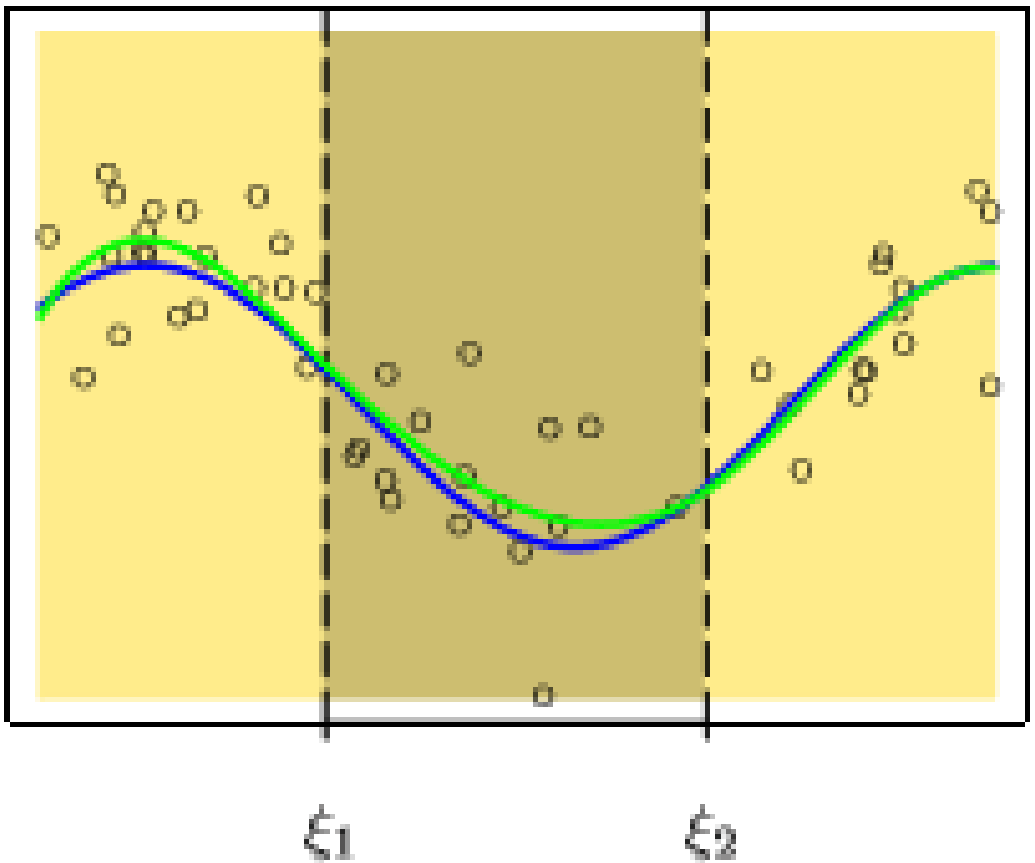
Por último, para generar un mejor suavizamiento, además de forzar la continuidad, se puede forzar la suavidad, utilizando restricciones al orden de las derivadas, como último caso si quisiéramos generar un polinomio cúbico a tramos con segundas derivadas continuas en los nodos, se tiene que la base tendría 12 parámetros - 6 restricciones, lo cual da la siguiente base:

$$h_1(X) = 1, \quad h_3(X) = X^2, \quad h_5(X) = (X - \xi_1)_+^3$$
$$h_2(X) = X, \quad h_3(X) = X^3, \quad h_6(X) = (X - \xi_2)_+^3$$

Este tipo de base se conoce como **spline cúbico**

Modelos Aditivos Generalizados

Polinomios a tramos



Modelos Aditivos Generalizados

Polinomios a tramos

En general cualquier base de un spline de orden M con nodos $\xi_j, \quad j = 1, \ldots, K$, con derivadas continuas hasta el orden $M-2$ puede se expresado como:

$$h_j(X) = X^{j-1}, \quad j = 1, \ldots, M$$
$$h_{M+t}(X) = (X - \xi_t)_+^{M-1}, \quad t = 1, \ldots, K.$$

Modelos Aditivos Generalizados

Polinomios a tramos

En general cualquier base de un spline de orden M con nodos $\xi_j, \quad j = 1, \ldots, K$, con derivadas continuas hasta el orden $M - 2$ puede se expresado como:

$$h_j(X) = X^{j-1}, \quad j = 1, \ldots, M$$
$$h_{M+t}(X) = (X - \xi_t)_+^{M-1}, \quad t = 1, \ldots, K.$$

En general no se debiera usar un spline de un orden más allá del cúbico.

Modelos Aditivos Generalizados

El problema que se tiene ahora es el de la selección de nodos. Claramente estos no se eligen a mano (a menos que se conozca muy bien la data). Supongamos que tenemos un conjunto maximal de nodos, esto normalmente es sinónimo de sobreajuste. Esto último se puede con regularización, es decir buscamos minimizar

$$\sum_{i=1}^{n}\left[y_i - \alpha\sum_{j=1}^{p} f(x_{ij})\right]^2 + \sum_{j=1}^{p}\lambda_j\int[f''(t)]^2 dt$$

El primer termino busca que la curva se acerque lo más posible a los datos, mientras que el segundo busca reducir en lo posible la curvatura de la función. Aquí el rol protagónico lo pasa a tomar λ, que define el trade-off entre ambas.

Modelos Aditivos Generalizados

Algoritmo: Retro ajuste

1. inicializar $\hat{\alpha} = \bar{y}, \quad \hat{f}_j = 0, \quad \forall i, j$
2. Iterar infinito hasta que $\hat{f}_j$ cambie menos que un umbral especificado.
3. Iterar sobre $j = 1, \ldots, p$

$$\hat{f}_j = \mathcal{S}_j \left[\{ y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ik}) \} \right]$$

$$\hat{f}_j = \hat{f}_j - \frac{1}{n} \sum_{i=1}^{n} \hat{f}_j(x_{ij})$$

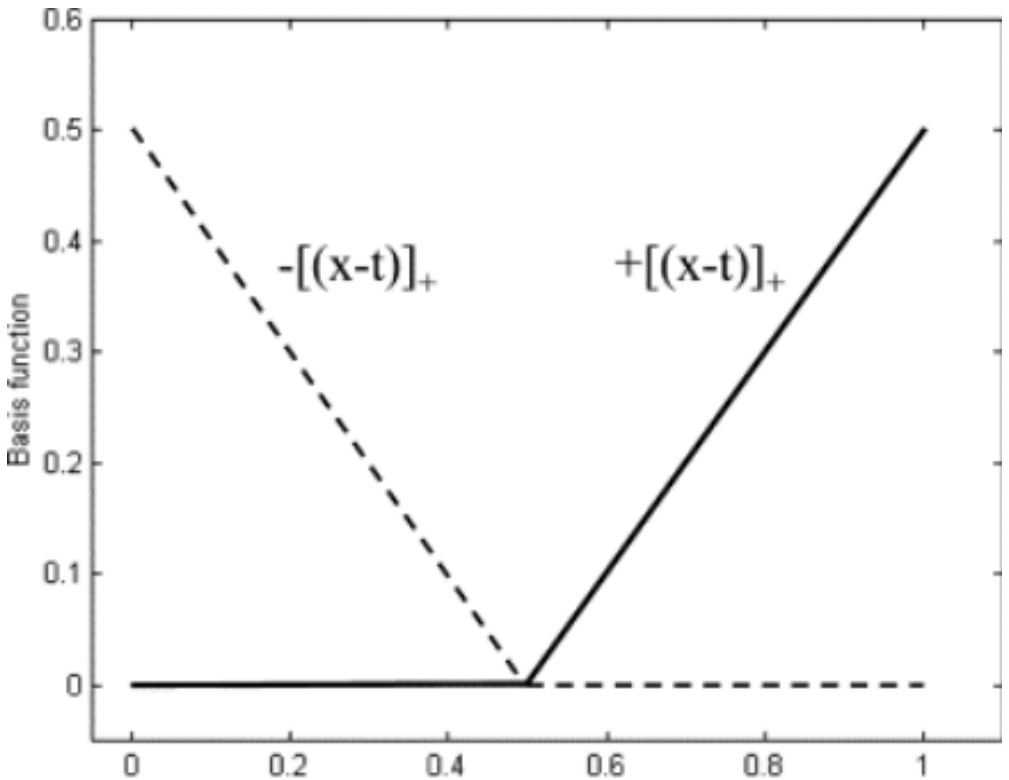
$\mathcal{S}_j$ Corresponde a una función de suavizamiento, normalmente splines cúbicos, pero también puede ser regresión local (LOESS) o suavizamiento Kernel.

Modelos Aditivos Generalizados

Ejemplo: Usando los datos de loan y las mismas variables anteriores, ajuste un modelo aditivo generalizado. Compare con una regresión logística simple su desempeño.

MARS

Multivariate Adaptative Regresion Splines (MARS) es una metodología que toma más valor al ser aplicada a problemas de alta dimensionalidad. Consiste en usar bases lineales de la forma $(x-t)_+$ y $(t-x)_+$



MARS

t correponde a un nodo, y la combinación de $(x-t)_+$ y $(t-x)_+$ es un par reflejado. La idea es armar pares reflejados para cada variable con nodos en cada punto observado, es decir

$$\mathcal{C} = \{(X_j - t)_+, (t - X_j)\}, \quad t \in \{x_{1j}, \ldots, x_{nj}\}, \quad j = 1, 2, \ldots, p$$

Claramente no se pueden usar todos los pares reflejados para el ajuste de un modelo. La estrategia es bastante simple, usar una selección Fordward, pero en lugar de las variables brutas, se van eligiendo entre los pares reflejados.