

Selección, Validación de Modelos y Riesgo Esperado

Febrero de 2023

Contenido

- 1 Selección, Validación de Modelos y Riesgo Esperado
- 2 Validación Cruzada
- 3 Bootstrap
- 4 Es el riesgo esperado una buena estimación del riesgo?
- 5 Validación de Modelos
 - Curva ROC I
 - Curva ROC II
 - Curva de calibración

Sesgo, Varianza y Riesgo Esperado

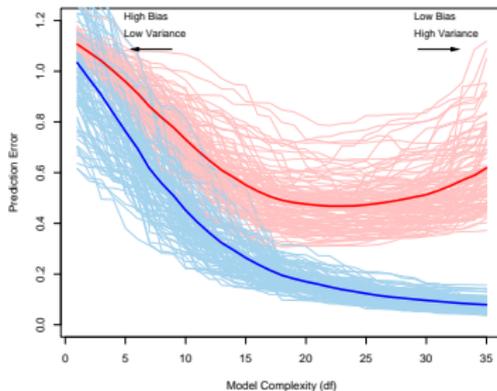


FIGURE 7.1. Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error $\overline{\text{err}}$, while the light red curves show the conditional test error $\text{Err}_{\mathcal{T}}$ for 100 training sets of size 50 each, as the model complexity is increased. The solid curves show the expected test error Err and the expected training error $\text{E}[\overline{\text{err}}]$.

- Elegir parámetros (meta parámetros).
- Estimar el error de prueba.

- El **error de entrenamiento** es:

$$\bar{err} = \frac{1}{N} \sum_{i=1}^n L(y_i, \hat{f}(x_i))$$

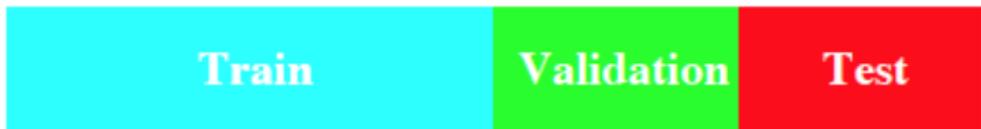
- El **error de prueba o generalización** es:

$$Err_{\tau} = E[L(X, \hat{f}(X)) | \tau]$$

- El **error esperado de prueba o generalización** es:

$$Err = E_{\tau}[E[L(X, \hat{f}(X)) | \tau]] = E[E_{\tau}]$$

- En ambientes ricos en datos podemos seleccionar modelos (metaparámetros) y estimar el riesgo.



- Si bien el mayor interés es la estimación del riesgo (condicional a una muestra), es en general muy difícil de hacerlo y lo máximo que se puede hacer es estimar el riesgo esperado.
- En esta parte vamos a discutir algunos métodos estadísticos para estimar el riesgo esperado. Algunos de estos métodos son analíticos (AIC, etc.) y otros basados en técnicas de remuestreo (validación cruzada y bootstrap).

Contenido

- 1 Selección, Validación de Modelos y Riesgo Esperado
- 2 Validación Cruzada**
- 3 Bootstrap
- 4 Es el riesgo esperado una buena estimación del riesgo?
- 5 Validación de Modelos
 - Curva ROC I
 - Curva ROC II
 - Curva de calibración

Validación Cruzada

- Cuando no se tiene tanta información se puede hacer validación cruzada.
- Esto permite estimar el riesgo esperado y seleccionar modelos.
- Validación cruzada de K muestras:
 - 1 Dividir en K muestras aleatorias la muestra original de tamaño N . Dada la muestra k se entrena el modelo sin los datos de esta muestra y se estima el error en esa muestra. El promedio de los errores es la estimación del riesgo esperado.
 - 2 Cuando $K = N$ se conoce como *leave out one cross validation*. En este caso el modelo estimado puede tener una varianza alta pero el sesgo en la estimación del error esperado es bajo.

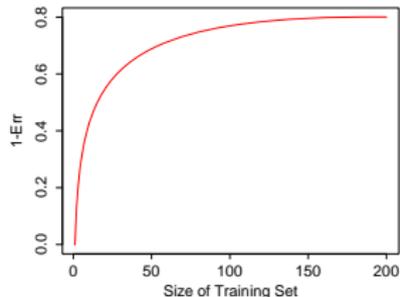


FIGURE 7.8. *Hypothetical learning curve for a classifier on a given task: a plot of $1 - \text{Err}$ versus the size of the training set N . With a dataset of 200 observations, 5-fold cross-validation would use training sets of size 160, which would behave much like the full set. However, with a dataset of 50 observations fivefold cross-validation would use training sets of size 40, and this would result in a considerable overestimate of prediction error.*

- El número de divisiones óptimo depende del número de datos. Un K alto con pocos datos sobre estima el error de prueba (varianza alta). Un K bajo subestima el error (sesgo alto).

Validación Cruzada: Uso adecuado

- Considere un problema con mucho predictores.
- Haga una selección de variables con base en la correlación (univariada) de las variables con la variable objetivo.
- Usando este subconjunto construya un clasificador.
- Use validación cruzada

Validación Cruzada: Uso adecuado

- Considere el siguiente ejemplo: una muestra de tamaño 50 con el mismo número de dos clases. 5000 predictores independientes de las clases. El mejor modelo tiene un error de clasificación del 50 %.
- Al usar la estrategia anterior eligiendo los 100 predictores que más correlación tienen con la variable objetivo y después usando un clasificador de un vecino más cercano se hace validación cruzada con $K = 5$.
- Si se simula 50 veces este procedimiento el error de validación cruzada es 3 %.

Validación Cruzada: Uso adecuado

- El problema es que la selección de variables se hace con todos los datos incluso aquellos que se van a utilizar para hacer la validación. Esto es un problema clásico de fuga de información (i.e, information leakage).
- En la muestra de validación la correlación es muy alta.

Validación Cruzada: Uso adecuado

- La forma correcta es: hacer la selección de variables como parte de la validación cruzada.
- En general cuando hay varias etapas en la metodología en cada remuestreo se debe aplicar la metodología.
- La única excepción es cuando se ha usado algún método para filtrar que no sea supervisado (no dependa de la variable objetivo). Por ejemplo, elegir aquellas variables con mayor varianza.

Validación Cruzada: Uso adecuado

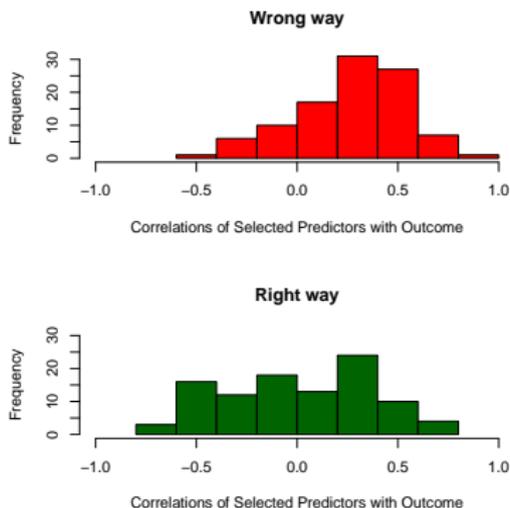


FIGURE 7.10. *Cross-validation the wrong and right way: histograms shows the correlation of class labels, in 10 randomly chosen samples, with the 100 predictors chosen using the incorrect (upper red) and correct (lower green) versions of cross-validation.*

Contenido

- 1 Selección, Validación de Modelos y Riesgo Esperado
- 2 Validación Cruzada
- 3 Bootstrap**
- 4 Es el riesgo esperado una buena estimación del riesgo?
- 5 Validación de Modelos
 - Curva ROC I
 - Curva ROC II
 - Curva de calibración

Bootstrap

- En general la metodología de bootstrap se utiliza para deducir las propiedades estadísticas de funciones de una muestra finitas (estimadores).
- Aquí vamos a explicar como usarlo para estimar el riesgo esperado.

Bootstrap: Estadísticos muestras finitas

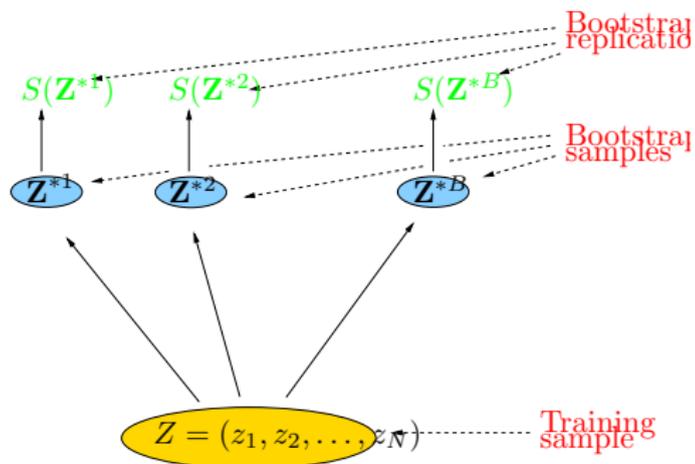


FIGURE 7.12. Schematic of the bootstrap process. We wish to assess the statistical accuracy of a quantity $S(\mathbf{Z})$ computed from our dataset. B training sets \mathbf{Z}^{*b} , $b = 1, \dots, B$ each of size N are drawn with replacement from the original dataset. The quantity of interest $S(\mathbf{Z})$ is computed from each bootstrap training set, and the values $S(\mathbf{Z}^{*1}), \dots, S(\mathbf{Z}^{*B})$ are used to assess the statistical accuracy of $S(\mathbf{Z})$.

Contenido

- 1 Selección, Validación de Modelos y Riesgo Esperado
- 2 Validación Cruzada
- 3 Bootstrap
- 4 Es el riesgo esperado una buena estimación del riesgo?
- 5 Validación de Modelos
 - Curva ROC I
 - Curva ROC II
 - Curva de calibración

Es el riesgo esperado una buena estimación del riesgo?

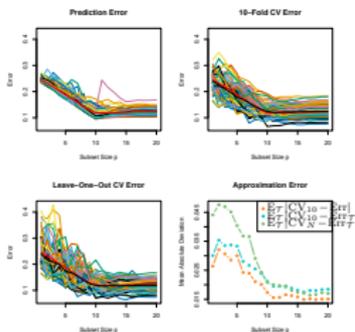


FIGURE 7.14. Conditional prediction-error $E_{T,T}$, 10-fold cross-validation, and leave-one-out cross-validation curves for a 100 simulations from the top-right panel in Figure 7.3. The thick red curve is the expected prediction error E_T , while the thick black curves are the expected CV curves $E_T CV_{10}$ and $E_T CV_N$. The lower-right panel shows the mean absolute deviation of the CV curves from the conditional error, $E_T |CV_K - E_{T,T}|$ for $K = 10$ (blue) and $K = N$ (green), as well as from the expected error $E_T |CV_{10} - E_T|$ (orange).

Contenido

- 1 Selección, Validación de Modelos y Riesgo Esperado
- 2 Validación Cruzada
- 3 Bootstrap
- 4 Es el riesgo esperado una buena estimación del riesgo?
- 5 Validación de Modelos
 - Curva ROC I
 - Curva ROC II
 - Curva de calibración

Validación de Modelos

- Modelos de regresión: AIC, R^2 , MAPE, etc.
- Modelo de clasificación: Curva ROC y curva de calibración

Curva ROC

- La curva ROC y el área bajo la curva es uno de los métodos más importantes de validación para problemas de clasificación.

3. Receiver operating characteristic curves

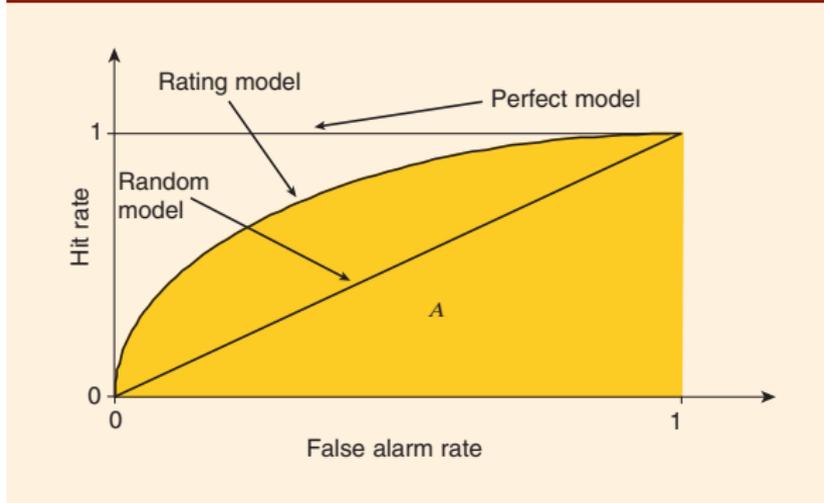


Figura: Hit rate: True Positive Rate. False alarm rate: False Positive Rate

- Los modelos de clasificación binaria se pueden extender a clasificación en múltiples categorías.

Curva ROC

- Por simplicidad definamos el score de un individuo como la probabilidad de incumplimiento.
- Considere una gráfica de la distribución acumulada del score de los buenos y los malos. El score que arroje la distancia máxima entre estas distribuciones es la distancia de Kolmogorov y Smirnov.
- Si representamos estas dos gráficas en un mismo gráfico se obtiene la curva ROC: en el eje y la función de distribución de los buenos y en el eje x la función de distribución de los malos: Sensibilidad (y) vs. (1-Especificidad (x)).
- La distancia KS corresponde al score donde la distancia horizontal entre la curva ROC y la diagonal es máxima (pendiente 1).
- El coeficiente de Gini es el doble del área entre la diagonal y la curva ROC.
- En el gráfico de la curva ROC, el KS corresponde al punto en el que la curva tiene pendiente 1 o mayor distancia a la diagonal.

Curva ROC

- Por simplicidad definamos el score de un individuo como la probabilidad de incumplimiento.
- Considere una gráfica de la distribución acumulada del score de los buenos y los malos. El score que arroje la distancia máxima entre estas distribuciones es la distancia de Kolmogorov y Smirnov.
- Si representamos estas dos gráficas en un mismo gráfico se obtiene la curva ROC: en el eje y la función de distribución de los buenos y en el eje x la función de distribución de los malos: Sensibilidad (y) vs. (1-Especificidad (x)).
- La distancia KS corresponde al score donde la distancia horizontal entre la curva ROC y la diagonal es máxima (pendiente 1).
- El coeficiente de Gini es el doble del área entre la diagonal y la curva ROC.
- En el gráfico de la curva ROC, el KS corresponde al punto en el que la curva tiene pendiente 1 o mayor distancia a la diagonal.

Curva ROC

- Por simplicidad definamos el score de un individuo como la probabilidad de incumplimiento.
- Considere una gráfica de la distribución acumulada del score de los buenos y los malos. El score que arroje la distancia máxima entre estas distribuciones es la distancia de Kolmogorov y Smirnov.
- Si representamos estas dos gráficas en un mismo gráfico se obtiene la curva ROC: en el eje y la función de distribución de los buenos y en el eje x la función de distribución de los malos: Sensibilidad (y) vs. (1-Especificidad (x)).
- La distancia KS corresponde al score donde la distancia horizontal entre la curva ROC y la diagonal es máxima (pendiente 1).
- El coeficiente de Gini es el doble del área entre la diagonal y la curva ROC.
- En el gráfico de la curva ROC, el KS corresponde al punto en el que la curva tiene pendiente 1 o mayor distancia a la diagonal.

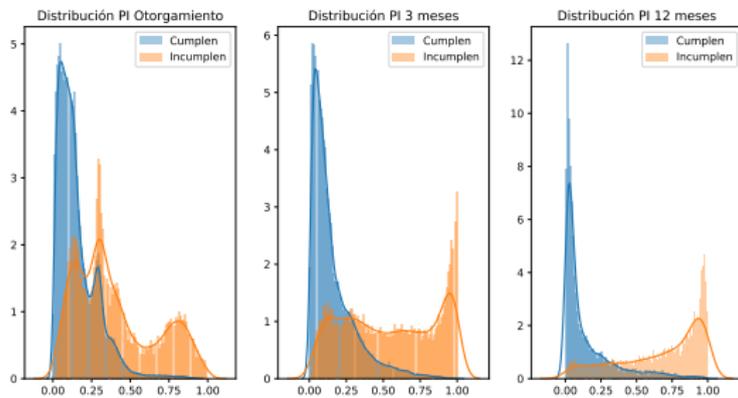
Curva ROC

- Por simplicidad definamos el score de un individuo como la probabilidad de incumplimiento.
- Considere una gráfica de la distribución acumulada del score de los buenos y los malos. El score que arroje la distancia máxima entre estas distribuciones es la distancia de Kolmogorov y Smirnov.
- Si representamos estas dos gráficas en un mismo gráfico se obtiene la curva ROC: en el eje y la función de distribución de los buenos y en el eje x la función de distribución de los malos: Sensibilidad (y) vs. (1-Especificidad (x)).
- La distancia KS corresponde al score donde la distancia horizontal entre la curva ROC y la diagonal es máxima (pendiente 1).
- El coeficiente de Gini es el doble del área entre la diagonal y la curva ROC.
- En el gráfico de la curva ROC, el KS corresponde al punto en el que la curva tiene pendiente 1 o mayor distancia a la diagonal.

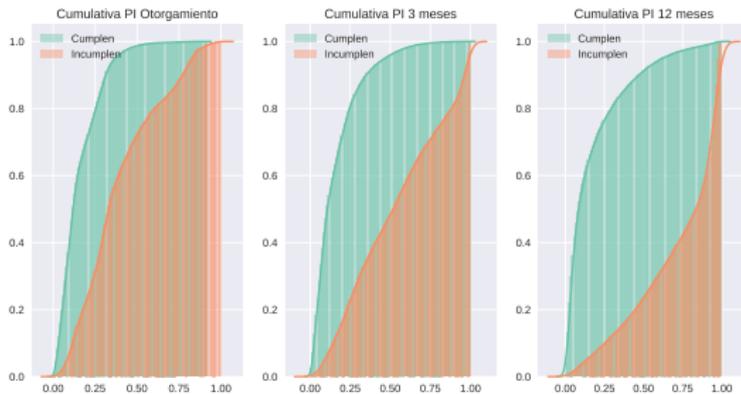
Curva ROC

- Por simplicidad definamos el score de un individuo como la probabilidad de incumplimiento.
- Considere una gráfica de la distribución acumulada del score de los buenos y los malos. El score que arroje la distancia máxima entre estas distribuciones es la distancia de Kolmogorov y Smirnov.
- Si representamos estas dos gráficas en un mismo gráfico se obtiene la curva ROC: en el eje y la función de distribución de los buenos y en el eje x la función de distribución de los malos: Sensibilidad (y) vs. (1-Especificidad (x)).
- La distancia KS corresponde al score donde la distancia horizontal entre la curva ROC y la diagonal es máxima (pendiente 1).
- El coeficiente de Gini es el doble del área entre la diagonal y la curva ROC.
- En el gráfico de la curva ROC, el KS corresponde al punto en el que la curva tiene pendiente 1 o mayor distancia a la diagonal.

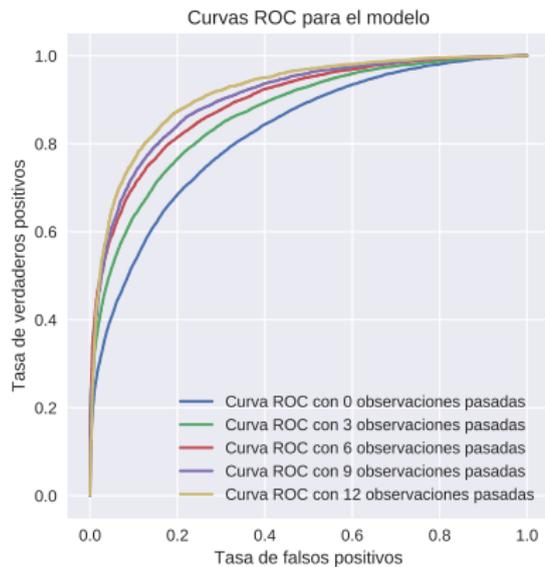
Curva ROC



Curva ROC



Curva ROC



Curva ROC: Explicación

- Es usual pensar en en términos de puntajes (score) en vez de PI.
- Definamos el score S de un individuo como $S = 1 - PI$.
- Dado un umbral $s \in [0, 1]$, definimos los clientes pronosticados positivos (i.e., clientes malos) como aquellos tales que $S \leq s$.
- Luego la distribución acumulada de los clientes positivos es $P_M(S \leq s) = 1 - F_M(1 - s)$.
- De esta forma se puede mostrar que las graficas anteriores son todas correctas si se cambia en el eje Y, por la distribución del score de los malos, y en el eje x, la distribución del score de los buenos.

Curva ROC: Explicación

- Sea M el total de unos, B el total de ceros y $T = M + B$.
- La tasa de verdaderos positivos (hit rate) se define como: Los que resultaron malos entre los que se pronosticaron malos sobre M .
- Los falsos positivos se define como: Los que salieron buenos entre los que se pronosticaron como malos sobre B .
- Entre más cerca este la curva del eje vertical y el eje horizontal superior, mejor el poder de clasificación global: para una misma tasa de falsos positivos, una mayor tasa de acierto.

Curva ROC: Explicación

- Sea M el total de unos, B el total de ceros y $T = M + B$.
- La tasa de verdaderos positivos (hit rate) se define como: Los que resultaron malos entre los que se pronosticaron malos sobre M .
- Los falsos positivos se define como: Los que salieron buenos entre los que se pronosticaron como malos sobre B .
- Entre más cerca este la curva del eje vertical y el eje horizontal superior, mejor el poder de clasificación global: para una misma tasa de falsos positivos, una mayor tasa de acierto.

Curva ROC: Explicación

- Sea M el total de unos, B el total de ceros y $T = M + B$.
- La tasa de verdaderos positivos (hit rate) se define como: Los que resultaron malos entre los que se pronosticaron malos sobre M .
- Los falsos positivos se define como: Los que salieron buenos entre los que se pronosticaron como malos sobre B .
- Entre más cerca este la curva del eje vertical y el eje horizontal superior, mejor el poder de clasificación global: para una misma tasa de falsos positivos, una mayor tasa de acierto.

Curva ROC: Explicación

- Sea M el total de unos, B el total de ceros y $T = M + B$.
- La tasa de verdaderos positivos (hit rate) se define como: Los que resultaron malos entre los que se pronosticaron malos sobre M .
- Los falsos positivos se define como: Los que salieron buenos entre los que se pronosticaron como malos sobre B .
- Entre más cerca este la curva del eje vertical y el eje horizontal superior, mejor el poder de clasificación global: para una misma tasa de falsos positivos, una mayor tasa de acierto.

Matriz de confusión

Obs\Pred	m	b
M	vp	fn
B	fp	vn

- Tasa de verdaderos positivos: $TVP = \frac{vp}{M}$. También se conoce como *recall*, *hit rate*, *sensitivity*.
- Tasa de falsos positivos: $TFP = \frac{fp}{B}$. También se conoce como *false alarm rate* (o *1-specificity*).
- Obsérvese que en la matriz de confusión TVP es la distribución del score de los malos evaluada en un umbral y TFP la distribución de los buenos evaluada en ese mismo umbral.

Curva ROC: Explicación

- El modelo aleatorio tiene como curva ROC la diagonal de 45 grados.
- Para ver esto fijemos un score s y supongamos que cuando tenemos un ejemplo con características x . El modelo aleatorio le asigna un score $f(x) \in [0, 1]$ que se distribuye uniforme y se marca como malo, 1 si $f(x) \leq s$.
- Entonces:

$$TVP = \frac{sT \times \frac{M}{T}}{M} = s$$

$$TFP = \frac{sT \times \frac{B}{T}}{B} = s$$

- Luego la curva ROC es la diagonal.

Curva ROC: Explicación

- El modelo aleatorio tiene como curva ROC la diagonal de 45 grados.
- Para ver esto fijemos un score s y supongamos que cuando tenemos un ejemplo con características x . El modelo aleatorio le asigna un score $f(x) \in [0, 1]$ que se distribuye uniforme y se marca como malo, 1 si $f(x) \leq s$.
- Entonces:

$$TVP = \frac{sT \times \frac{M}{T}}{M} = s$$

$$TFP = \frac{sT \times \frac{B}{T}}{B} = s$$

- Luego la curva ROC es la diagonal.

Curva ROC: Explicación

- El modelo aleatorio tiene como curva ROC la diagonal de 45 grados.
- Para ver esto fijemos un score s y supongamos que cuando tenemos un ejemplo con características x . El modelo aleatorio le asigna un score $f(x) \in [0, 1]$ que se distribuye uniforme y se marca como malo, 1 si $f(x) \leq s$.
- Entonces:

$$TVP = \frac{sT \times \frac{M}{T}}{M} = s$$

$$TFP = \frac{sT \times \frac{B}{T}}{B} = s$$

- Luego la curva ROC es la diagonal.

Curva ROC: Explicación

- El modelo aleatorio tiene como curva ROC la diagonal de 45 grados.
- Para ver esto fijemos un score s y supongamos que cuando tenemos un ejemplo con características x . El modelo aleatorio le asigna un score $f(x) \in [0, 1]$ que se distribuye uniforme y se marca como malo, 1 si $f(x) \leq s$.
- Entonces:

$$TVP = \frac{sT \times \frac{M}{T}}{M} = s$$

$$TFP = \frac{sT \times \frac{B}{T}}{B} = s$$

- Luego la curva ROC es la diagonal.

Curva ROC: Explicación

- El modelo aleatorio tiene como curva ROC la diagonal de 45 grados.
- Para ver esto fijemos un score s y supongamos que cuando tenemos un ejemplo con características x . El modelo aleatorio le asigna un score $f(x) \in [0, 1]$ que se distribuye uniforme y se marca como malo, 1 si $f(x) \leq s$.
- Entonces:

$$TVP = \frac{sT \times \frac{M}{T}}{M} = s$$

$$TFP = \frac{sT \times \frac{B}{T}}{B} = s$$

- Luego la curva ROC es la diagonal.

Curva de calibración

- Mide el error entre las frecuencias pronosticadas de un evento y las frecuencias observadas de ocurrencia.
- En las aplicaciones de aprendizaje de máquinas, a medicina es estándar y se usa una prueba χ^2 para determinar la significancia estadística de la diferencia.
- Es manipulable.

Curva de calibración

- Mide el error entre las frecuencias pronosticadas de un evento y las frecuencias observadas de ocurrencia.
- En las aplicaciones de aprendizaje de máquinas, a medicina es estándar y se usa una prueba χ^2 para determinar la significancia estadística de la diferencia.
- Es manipulable.

Curva de calibración

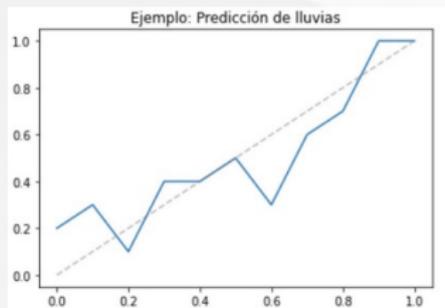
- Mide el error entre las frecuencias pronosticadas de un evento y las frecuencias observadas de ocurrencia.
- En las aplicaciones de aprendizaje de máquinas, a medicina es estándar y se usa una prueba χ^2 para determinar la significancia estadística de la diferencia.
- Es manipulable.

Curvas de calibración

Cuando nuestras predicciones son probabilidades $p \in [0, 1]$, es importante verificar que las probabilidades reflejen la realidad. Cuando esto ocurre, decimos que nuestro modelo está **calibrado**.

Ejemplo

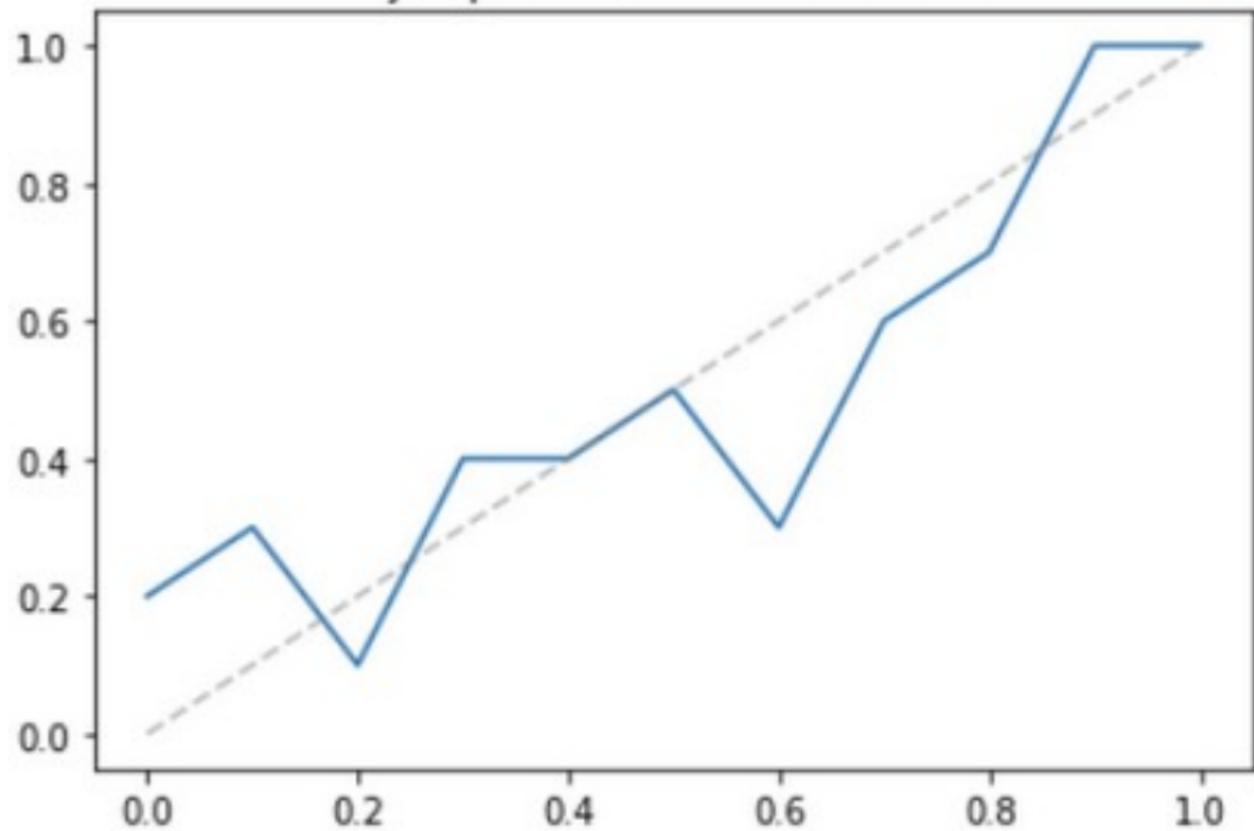
Quiero predecir si va a llover o no en Bogotá. Para esto, obtengo la probabilidad de que llueva para 110 días en total. Este es el resumen de los resultados:



Probabilidad asignada	Número de días	Número de días donde realmente llovió	Porcentaje de días en que llovió
0	10	2	0.2
0.1	10	3	0.3
0.2	10	1	0.1
0.3	10	4	0.4
0.4	10	4	0.4
0.5	10	5	0.5
0.6	10	3	0.3
0.7	10	6	0.6
0.8	10	7	0.7
0.9	10	10	1
1	10	10	1

Curva de Calibración: Ejemplo

Ejemplo: Predicción de Lluvias



Curva de Calibración: Condicional

Curva de calibración para genero

