



UNIVERSIDAD DE CUENCA
desde 1867

Evaluación de la Hipótesis

Andres Auquilla

2024

Content



Comparar modelos

Técnicas para comparar calidad de soluciones

Métricas de Clasificación

Técnicas clásicas e interpretación

Métricas de Regresión

Tipos de métricas e interpretación

Content



Comparar modelos

Técnicas para comparar calidad de soluciones

Métricas de Clasificación

Técnicas clásicas e interpretación

Métricas de Regresión

Tipos de métricas e interpretación

La **precisión de un modelo** es la probabilidad que la hipótesis prediga correctamente una instancia aleatoria de la población

$$Acc(h) = P[h(X) = c(X)] = 1 - error(h)$$

- Estimar precisión = estadísticas estándar
- Distribuciones binomiales y normales
- Intervalos de confianza, pruebas de hipótesis

Distribuciones binomiales

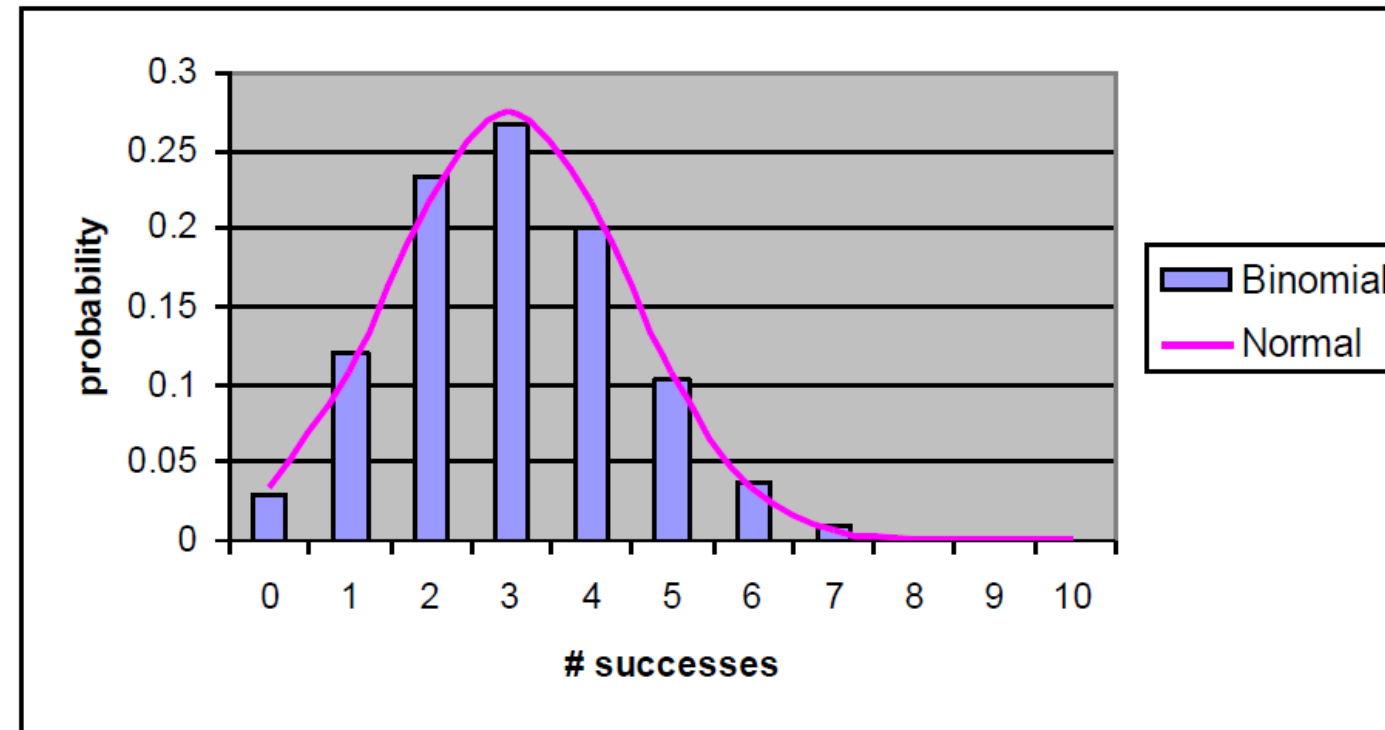
Un experimento exitoso con probabilidad p se repite n veces. ¿Cuál es la probabilidad de tener x intentos exitosos?

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

- Se asume: p constante y experimentos independientes
- Dado un h con accuracy p , n instancias en el test set, se calcula la probabilidad de hacer x predicciones correctas en este test set

Las distribuciones binomiales pueden ser aproximadas por una distribución normal

Example:
 $n=10$, $p=0.3$



Esta aproximación es muy utilizada en la práctica

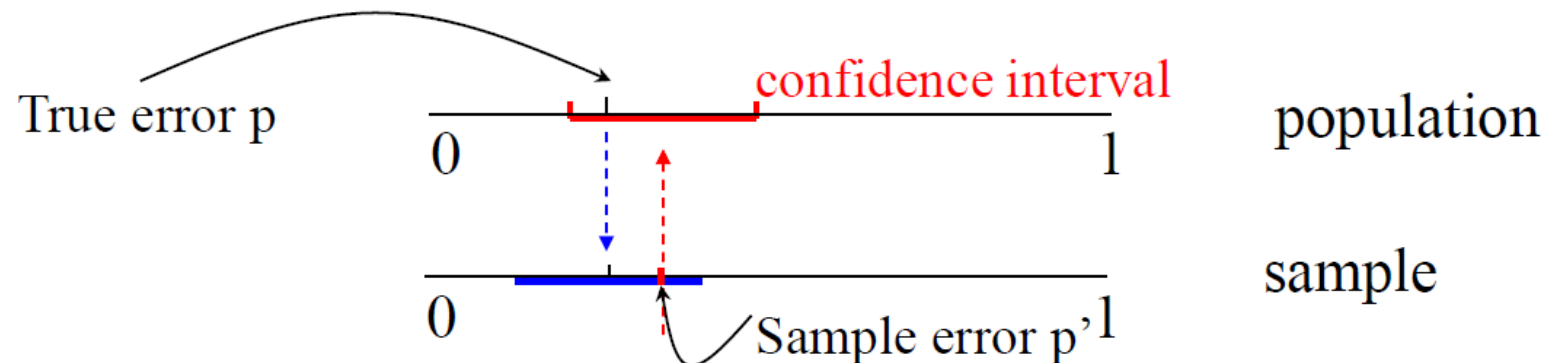
Intervalos de confianza

Con p y n , se puede calcular un intervalo en el cuál x (o x/n) se encuentra con probabilidad cercana a 1

- Con una distribución binomial o normal

En práctica, se quiere hacer lo opuesto

- Dado $p' = x/n$, proveer un intervalo para p que contenga p' con una probabilidad cercana a 1
- El intervalo que contiene p con probabilidad c se denomina intervalo de confianza



Una **prueba de hipótesis** implica probar H_0 mediante el análisis de una muestra

Principio de la prueba de hipótesis

- H_0 se prueba en base a una muestra
- Si la muestra indica que $H_0 = \text{true}$ es improbable, se rechaza la hipótesis H_0

Ejemplo:

- Se asume: h predice correctamente en el 90% de los casos ($H_0: p = 0.9$)
- En el test set se obtiene $p' = 0.8$
- H_0 se rechaza si el intervalo de confianza alrededor de p' no contiene p

Las pruebas estadísticas intentan probar hipótesis mediante el cálculo de un valor: **p-value**

Problema: Se reclutan 10 personas con la finalidad de probar si existe una relación entre el ejercicio y pérdida de peso

Se colectaron los datos y se obtiene

- Promedio de pérdida de peso: 2 kg
- Desviación estándar en la muestra: 1 kg

Experimento

- H_0 : el ejercicio no afecta la pérdida de peso ($\mu = 0$)
- H_A : el ejercicio promueve la pérdida de peso ($\mu > 0$)

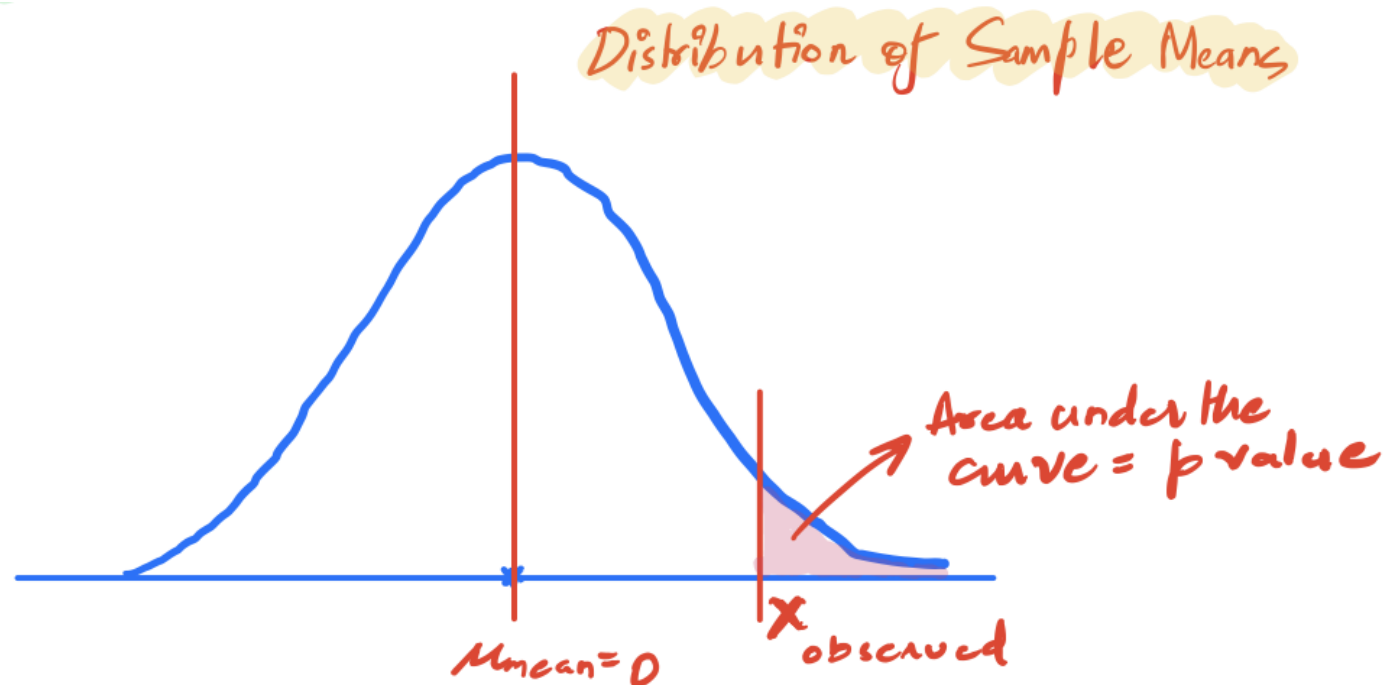
Asumiendo que H_0 es cierta, cuál es la probabilidad de observar un promedio en la población de 2 o más kilogramos

Si esta probabilidad es menor a un valor de threshold, H_0 se rechaza

Esta probabilidad es el p-value. El threshold se denomina nivel de significancia α (típicamente $\alpha = 0.05$)

Asumiendo que H_0 es cierto, ¿cual es la probabilidad de obtener un valor de 2 kg o más?

Teorema del límite central: población (μ, σ) ,
muestra $(\mu, \sigma/\sqrt{n_{\text{muestra}}})$



El valor de p se calcula:

```
from scipy.stats import norm
import numpy as np

p = 1-norm.cdf(2, loc=0, scale = 1/np.sqrt(10))
print(p)
-----
1.269814253745949e-10
```

$p < \alpha$ por ende, se rechaza H_0

Los test se asumen que son muestras aleatorias e independientes

La hipótesis h no es independiente del set de entrenamiento

Si se denota lo siguiente

- $error_{Tr}(h)$: error de h en el training set
- $error(h)$: error de h en la población
- $error_{Te}(h)$: error de h en el test set

Entonces, $E()$ denota el valor esperado

- $E(error_{Tr}(h)) < error(h)$: overfitting
- $E(error_{Tr}(h)) > error(h)$: bias del estimador $error_{Tr}$
- $E(error_{Te}(h)) = error(h)$: estimador sin sesgo

¿Cómo se pueden crear test de prueba independientes?

Lo más simple: dividir el dataset original

2/3 para entrenamiento y 1/3 para prueba

Cuando no hay muchos datos disponibles

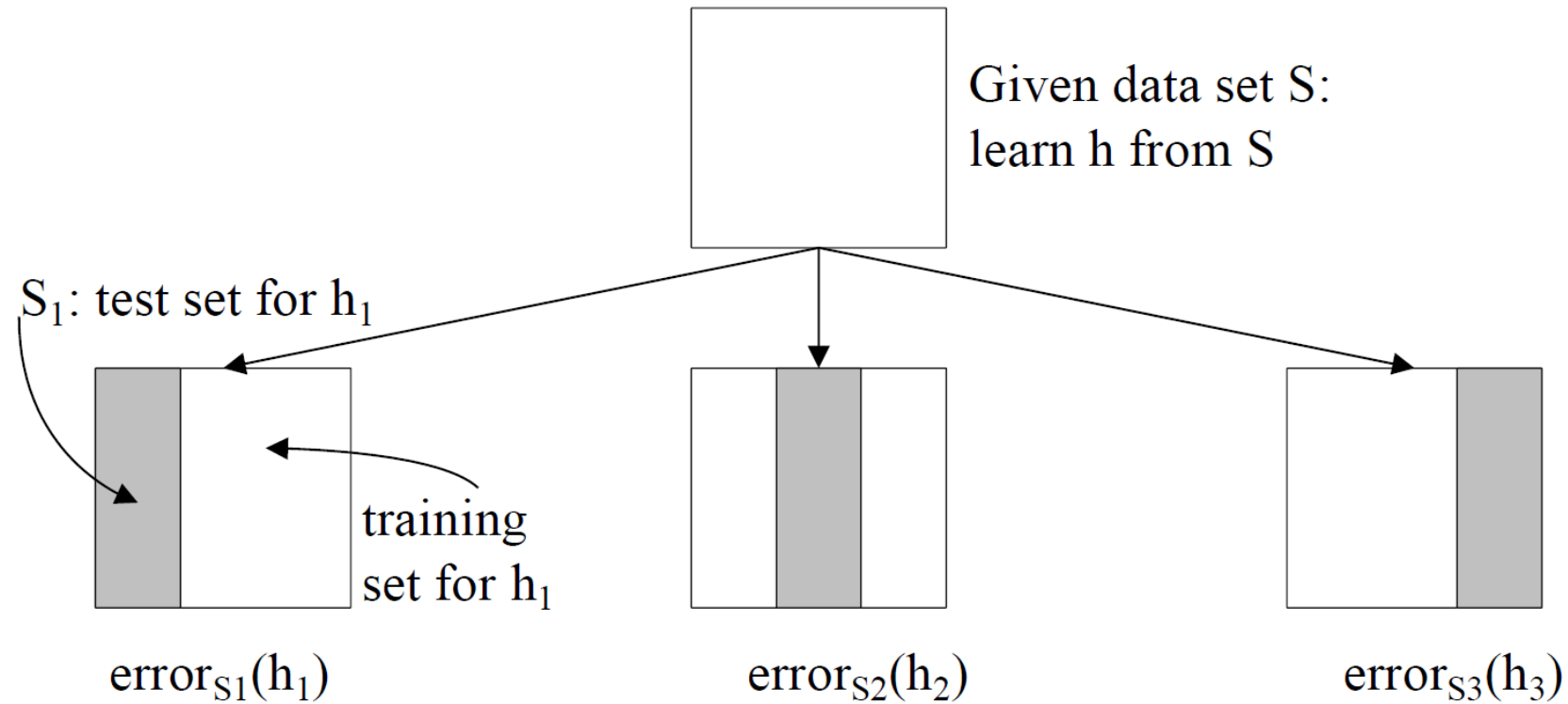
Training sets pequeños: más difícil entrenamiento, menos precisión

Otra solución popular: cross-validation

Dataset se parte n veces para entrenar hipótesis h_i

El dataset S se divide en varias particiones (folds):

k-fold cross-validation



$$error_S(h) = \sum_{i=1}^f \frac{|S_i|}{|S|} error_{S_i}(h_i)$$

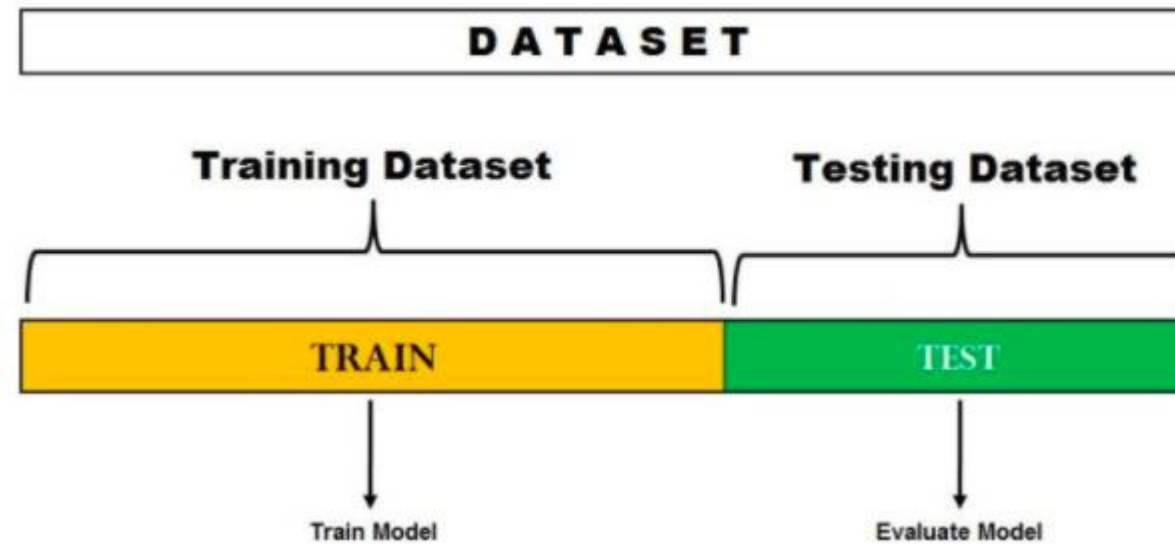
3-fold cross-validation

Existen varios tipos de cross-validation que pueden ser útiles en determinados casos

Tipos de cross-validation

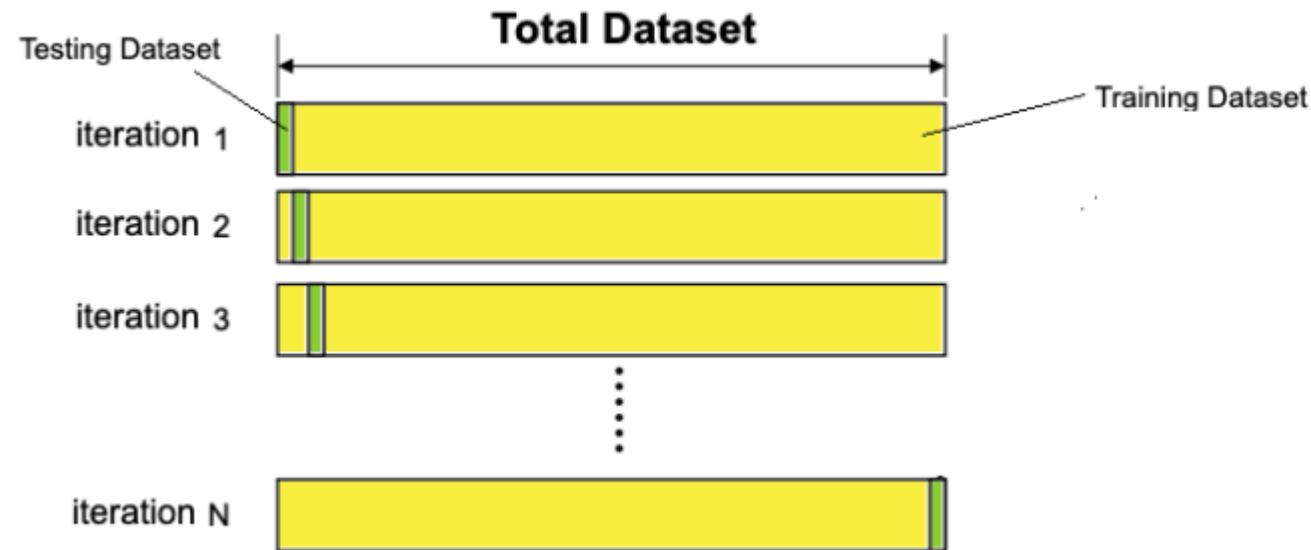
- Hold-out validation
- Leave-one-out cross validation (LOOCV)
- K-fold cross validation
- Stratified K-fold cross validation
- Cross validation for time series

Hold-out validation es la forma tradicional de partir los datos



Útil cuando hay gran cantidad de datos. Rápido, pero podrían ocultarse patrones importantes en el test set

Leave-One-Out se utiliza mayormente cuando hay muy pocos datos de entrenamiento



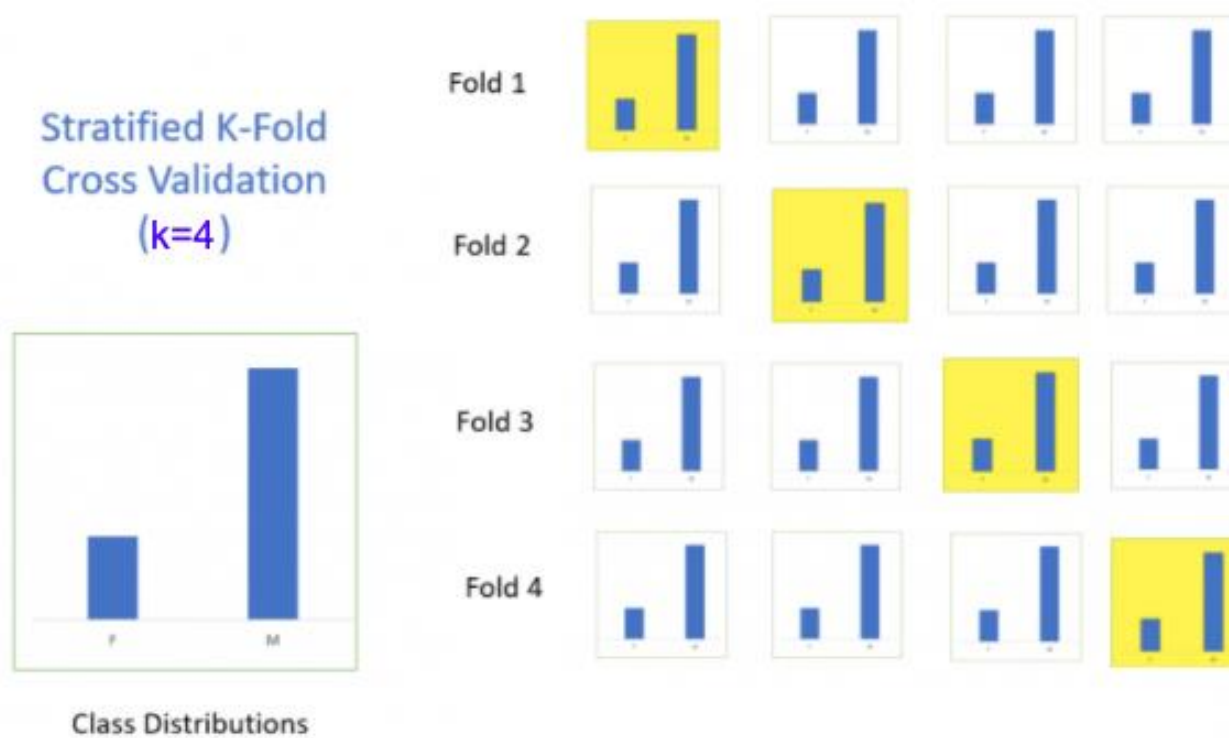
Bajo Bias ya que cubre todas las instancias. Proceso lento cuando existe gran cantidad de elementos

K-fold cross validation divide al dataset en K particiones aleatorias donde una de ella se usa para probar el modelo

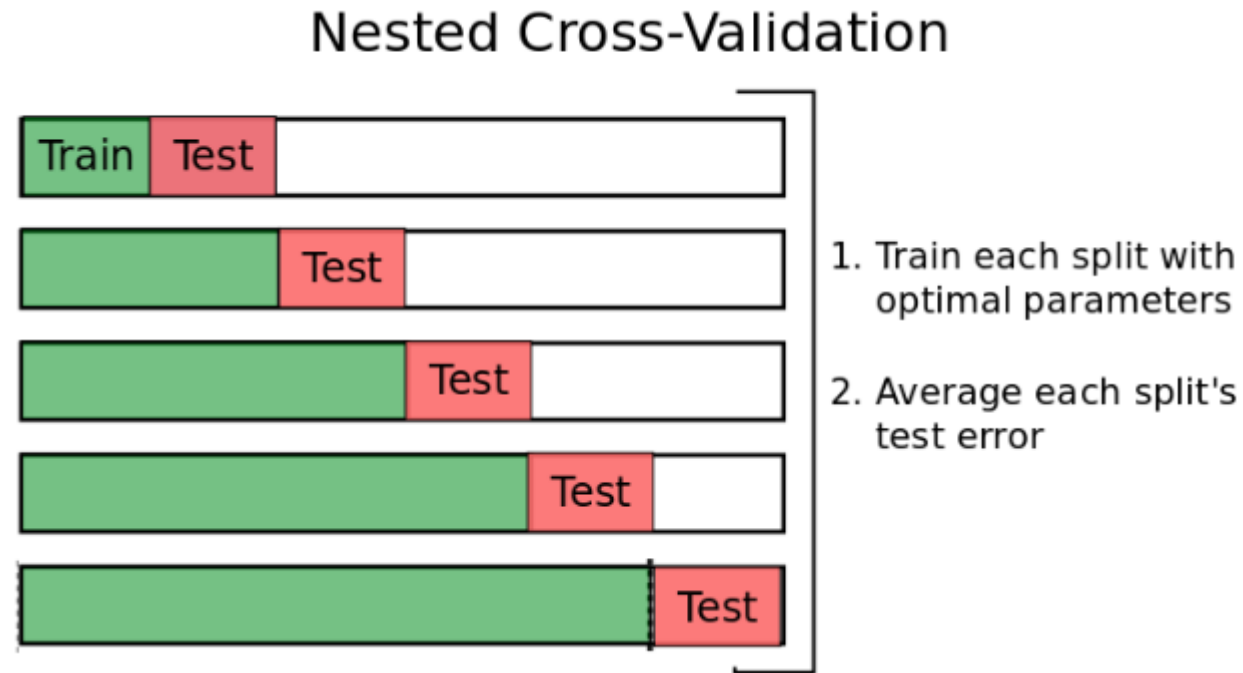


Funciona bien con datos que no son secuenciales

Stratified K-fold CV mantiene particiones de datos con la misma relación de frecuencia de clases



Cross validation para series de tiempo



Dada la naturaleza secuencia de los datos, este orden se preserva y se parten los datos en varias particiones

Cross-validation no es completamente “sin sesgo”

Usualmente tiene un pequeño “bias pesimista”

- S contiene más elementos que $S - S_i$
- Diferentes folds no son completamente independientes:
e.j. 10-fold cv con datos 5+ y 5-

Aún así, es preferible usarlo en vez de un solo accuracy sobre el train set

Dadas dos hipótesis, ¿Cuál de ellas tiene el menor error?

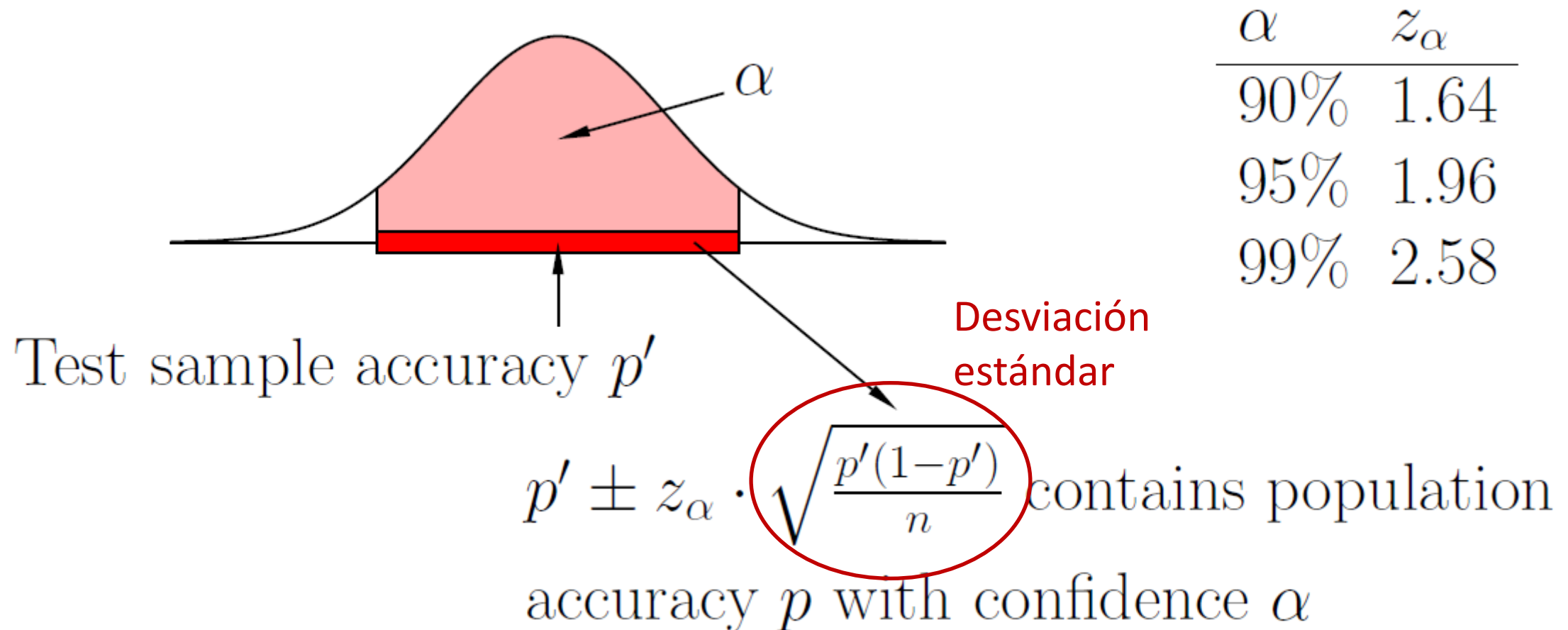
Prueba de hipótesis estadística

- Hipótesis: ambos son equivalentes
- Si la hipótesis es rechazada, se acepta que hay uno mejor

Dos casos:

- Comparar 2 hipótesis usando diferentes test sets
- Comparar 2 hipótesis usando el mismo test set

Un **intervalo de confianza** del accuracy/error de una hipótesis puede establecerse mediante una distribución normal



Comparar dos hipótesis (de forma general)

Para comparar h_1 y h_2 , se estima $p_1 - p_2$ de muestras $S_1(p'_1)$ y $S_2(p'_2)$

- Si $p_1 - p_2 > 0$: h_1 es mejor (el intervalo de confianza está a la derecha de 0)
- De forma similar, < 0 : h_2 es mejor
- De otra manera, no existen diferencias

Fórmula para el intervalo de confianza de la diferencia:

$$p'_1 - p'_2 \pm z_\alpha \sqrt{\frac{p'_1(1 - p'_1)}{n_1} + \frac{p'_2(1 - p'_2)}{n_2}}$$

Comparando 2 hipótesis en el mismo dataset

Cuando se comparan 2 hipótesis sobre el mismo test set

- Hay métodos más poderosos disponibles
- Usa más información del test set
- La influencia de ejemplos fáciles/difíciles se remueve

Un método más informativo

- Para cada ejemplo, comparar h_1 y h_2 cuan frecuente fue que h_1 fue correcto y h_2 incorrecto en el mismo ejemplo, vs el caso contrario
- McNemar test

El test de **McNemar** es la prueba estadística más comúnmente utilizada para comparar 2 hipótesis

Considere la siguiente tabla

| | h1 correct | h1 wrong |
|------------|------------|----------|
| h2 correct | A | B |
| h2 wrong | C | D |

Si h_1 es equivalente a h_2 :

- $P(C) = P(B) = 0.5$
- Por lo tanto, se espera que $B \approx C \approx (B + C)/2$
- B y C siguen una distribución binomial

El test de **McNemar** es la prueba estadística más comúnmente utilizada para comparar 2 hipótesis

Considere la siguiente tabla

| | h1 correct | h1 wrong |
|------------|------------|----------|
| h2 correct | A | B |
| h2 wrong | C | D |

Rechazar la equivalencia si B se desvía mucho de $(B + C)/2$

Ejemplo de comparación

| | h1 correct | h1 wrong |
|------------|------------|----------|
| h2 correct | 45 | 10 |
| h2 wrong | 0 | 45 |

Con tests sets independientes

- 55-45 en favor de h_2
- Resultados no muy convincentes

Comparación sobre el mismo test set

- Resultados más convincentes: 10-0 en favor de h_2

Ejemplo de comparación

| | h1 correct | h1 wrong |
|------------|------------|----------|
| h2 correct | 45 | 10 |
| h2 wrong | 0 | 45 |

h2 es claramente mejor que h1

Podría no ser descubierto usando una comparación “conservadora”

Se debe calcular un estadístico p para aceptar o rechazar la hipótesis nula $H_0: P(B) = P(C)$

H_0 implica que ambas hipótesis son equivalentes

Cuando $B + C < 25$, p se calcula como sigue (McNemar exacto)

- $p = 2 \sum_{i=B}^n \binom{n}{i} 0.5^i (1 - 0.5)^{n-i}$
- Donde $n = B + C$
- Esta formulación asume que $B \geq C$. Caso contrario, en la sumatoria cambia $i = C$
- Si $p < \alpha$, donde α representa el nivel de significancia, e.j. $\alpha = 0.05$, se rechaza H_0

Cuando $B - C \geq 25$, se debe calcular una estadística χ^2

Cuando $B + C \geq 25$, se calcula el valor de χ^2

- $\chi^2 = \frac{(B-C)^2}{(B+C)^2}$
- χ^2 sigue una distribución chi-cuadrada con un grado de libertad
- Luego de determinar un nivel de significancia, e.j. $\alpha = 0.05$, se calcula un valor de p
- p representa la probabilidad de que el valor de χ^2 pertenezca a esta distribución

Modelos pueden ser comparados a través la prueba estadística de hipótesis

Q1: dadas hipótesis h_1 y h_2 , cuál tiene mejor accuracy predictivo

Q2: dados los modelos L_1 y L_2 , y un dataset S , cual clasificador puede construir la mejor hipótesis de S

- Las hipótesis pueden variar
- Más difícil de responder que Q1

Modelos pueden ser comparados a través la prueba estadística de hipótesis

Para varios S_i , similares a S

- Dividir S_i en entrenamiento S_{tr} y prueba S_{te}
- Aprender h_1 y h_2 de S_{tr} usando L_1 y L_2
- Calcular el error $\delta_i = error_{S_{tr}}(h_1) - error_{S_{te}}(h_2)$

Prueba de hipótesis/intervalo de confianza la media de δ

¿Qué sucede si hay un número de datos limitados?

Si la cantidad de datos es limitada, se pueden dividir varias veces al mismo dataset

Repetir pruebas sobre el mismo dataset

- Cross validation: n splits en S_{tr} y S_{te}
- 5 cv, 10 cv, etc.

Problema: existencia de dependencias entre los data sets

- Puede generar inferencias equivocadas
- Alta probabilidad de error tipo I: inferir que un modelo es mejor que otro cuando no lo es

Una posible solución: 5 veces 2 cv

- Dietterich (1998)
- La mejor solución: recolectar más datos

5*2 CV Paired t Test se utiliza cuando se puede dividir al dataset multiples veces

Repetir 5 veces 2-fold CV

- S_1 y S_2 de igual tamaño. Cada algoritmo se entrena con uno y se prueba con otro

Se producen cuatro estimaciones de error $p_A^{(1)}$ y $p_B^{(1)}$ (entrenados en S_1 y probados en S_2), y $p_A^{(2)}$ y $p_B^{(2)}$

Se restan las diferencias de errores $p^{(1)} = p_A^{(1)} - p_B^{(1)}$ y $p^{(2)} = p_A^{(2)} - p_B^{(2)}$

Se estima la varianza $s^2 = (p^{(1)} - \bar{p})^2$ donde $\bar{p} = (p^{(1)} + p^{(2)})/2$

5*2 CV Paired t Test se utiliza cuando se puede dividir al dataset multiples veces

S_i^2 es la varianza calculada en la i_{th} iteración, $p_1^{(1)}$ es la $p^{(1)}$ de la primera de las 5 repeticiones. Se define la siguiente estadística

$$\blacksquare \quad \tilde{t} = \frac{p_1^{(1)}}{\sqrt{\frac{1}{5} \sum_{i=1}^5 S_i^2}}$$

Usando esta estadística, el valor de p puede ser calculado y comparado con el valor de significancia $\alpha = 0.05$. Si $p > \alpha$ se rechaza la hipótesis nula: existe diferencia entre los modelos

De esta forma se reduce la probabilidad de obtener un error de tipo 1 (determinar que un clasificador es mejor que otro, cuando no lo es)

K-fold cross-validated paired t test es otro método para comparar modelos

Los sets de entrenamiento se superponen, también no existe completa independencia entre los sets.

No se recomienda su uso en la práctica

Content



Comparar modelos

Técnicas para comparar calidad de soluciones

Métricas de Clasificación

Técnicas clásicas e interpretación

Métricas de Regresión

Tipos de métricas e interpretación





Muchas medidas de calidad se basan en la **matriz de confusión**

Para un problema con 2 clases, se genera una tabla con 4 valores

| | | Actual Values | |
|------------------|--------------|---------------|--------------|
| | | Positive (1) | Negative (0) |
| Predicted Values | Positive (1) | TP | FP |
| | Negative (0) | FN | TN |

Mediante esta table se pueden calcular métricas como: Recall, Precision, Specificity, Accuracy

Los valores de la matriz de confusión deben ser interpretados con cuidado

| | | Actual Values | |
|------------------|---|--|--|
| | | 1 | 0 |
| Predicted Values | 1 | TRUE POSITIVE  | FALSE POSITIVE  TYPE 1 ERROR |
| | 0 | FALSE NEGATIVE  TYPE 2 ERROR | TRUE NEGATIVE  |

Type 1 Error

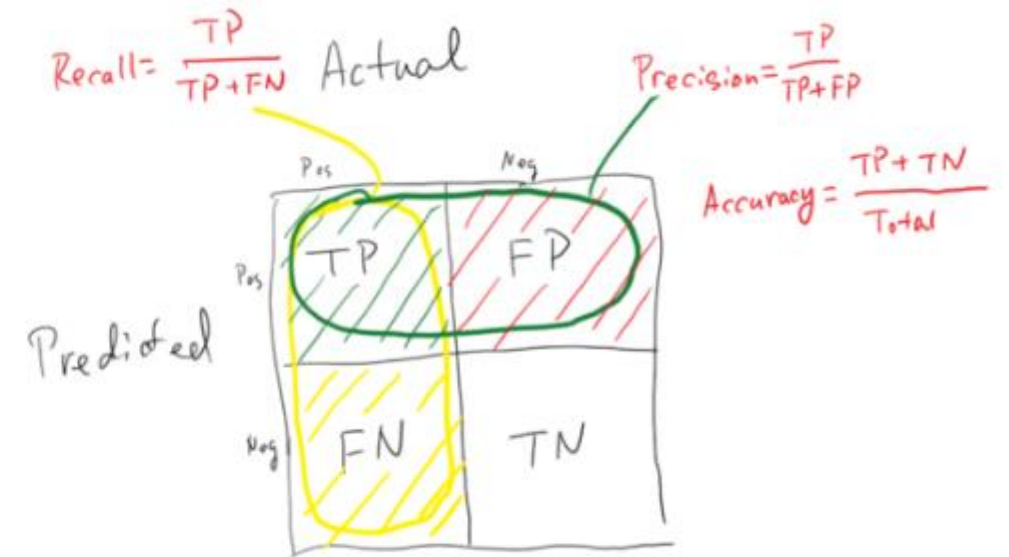
- El modelo predijo positivo cuando era falso

Type 2 Error

- El modelo predijo negativo cuando era positivo

Se pueden derivar múltiples métricas de una matriz de confusión

| y | y pred | output for threshold 0.6 | Recall | Precision | Accuracy |
|---|--------|--------------------------|------------|------------|------------|
| 0 | 0.5 | 0 | 1/2 | 2/3 | 4/7 |
| 1 | 0.9 | 1 | | | |
| 0 | 0.7 | 1 | | | |
| 1 | 0.7 | 1 | | | |
| 1 | 0.3 | 0 | | | |
| 0 | 0.4 | 0 | | | |
| 1 | 0.5 | 0 | | | |



Dependiendo del problema, otras medidas de calidad deberían ser analizadas

Recall (también conocido como Sensitivity o True Positive Rate)

- De todos los casos positivos, cuantos fueron clasificados correctamente
- La habilidad del modelo de encontrar los casos relevantes
- $Recall = \frac{T_P}{T_P + F_N}$

Precision

- De todos los elementos positivos predichos, cuantos son realmente positivos
- $Presicion = \frac{T_P}{T_P + F_P}$

Dependiendo del problema, otras medidas de calidad deberían ser analizadas

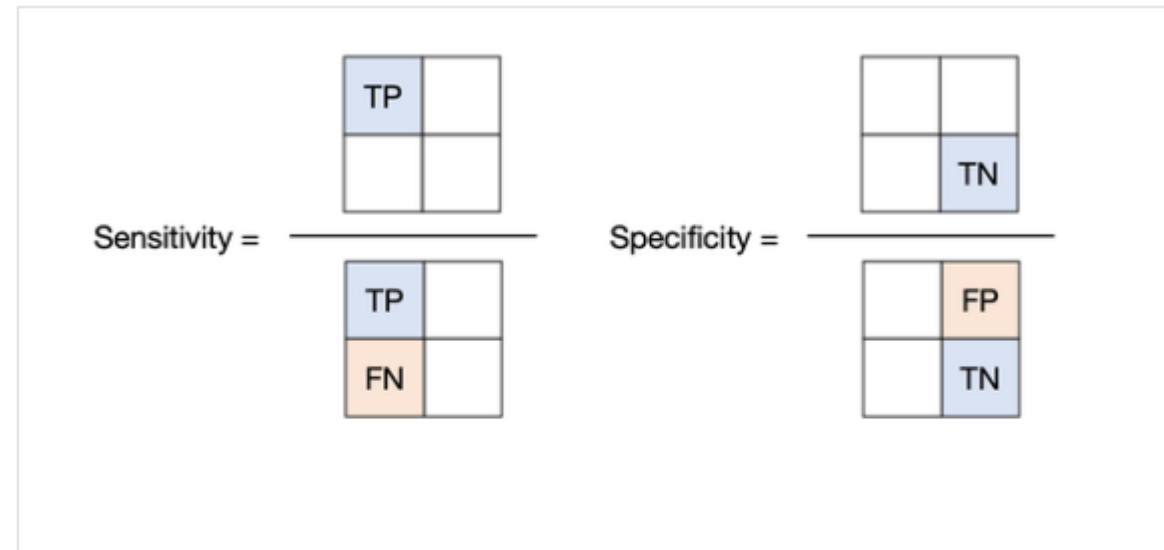
Accuracy

- De todos los casos positivos, cuantos fueron clasificados correctamente
- $Acc = \frac{T_P + T_N}{N}$

F1

- Promedio ponderado entre precisión y recall (media armónica)
- $Presicion = \frac{2 \times Recall \times Presicion}{Recall + Presicion}$

Cuando existe un desbalance de clases en el dataset, se puede utilizar **Balanced Accuracy**



Balanced accuracy is simply the arithmetic mean of the two:

$$\text{Balanced accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}.$$

| | Actual positive (1) | Actual negative (0) |
|------------------------|---------------------|---------------------|
| Predicted positive (1) | 5 | 50 |
| Predicted negative (0) | 10 | 10000 |

The accuracy of this classifier, i.e. the proportion of correct predictions, is

$$\frac{5 + 10000}{5 + 50 + 10 + 10000} \approx 99.4\%.$$

That sounds really impressive until you realize that simply by predicting all negative, we would have obtained an accuracy of

$$\frac{0 + 10050}{0 + 0 + 15 + 10050} \approx 99.9\%,$$

which is better than our classifier!

Balanced accuracy attempts to account for the imbalance in classes. Here is the computation for balanced accuracy for our classifier:

$$\begin{aligned} \text{Sensitivity} &= \frac{5}{5 + 10} \approx 33.3\%, \\ \text{Specificity} &= \frac{10000}{50 + 10000} \approx 99.5\%, \\ \text{Balanced accuracy} &= \frac{\text{Sensitivity} + \text{Specificity}}{2} \\ &\approx \frac{33.3\% + 99.5\%}{2} \\ &= 66.4\%. \end{aligned}$$

Existen ocasiones donde la evaluación a través del accuracy no es apropiado

Accuracy puede ser engañoso

Cuando las clases no están balanceadas

Accuracy se vuelve inestable

Asume costos de clasificación incorrecta simétricos

Alternativas

Correlación, ROC

La métrica accuracy puede ser engañosa

Accuracy de 99% ¿Es un buen score?

- Si -> cuando 50% “+” y 50% “-”
- No -> cuando 1% “+” y 99% “-”

Accuracy es una medida relativa

- Debería ser comparada contra una medida en donde siempre se elija la clase con más soporte
- Accuracy base $\max\{a + c, b + d\}/T$

Incluso así, puede ser engañosa

Una alternativa al Accuracy es la **correlación**

$$\text{Correlación } \varphi = \frac{ad-bc}{\sqrt{T_{pos}T_{neg}T_{+}T_{-}}}$$

- Cercano a 1: alta correlación entre predicciones y clases
- Cercano a 0: no hay correlación
- Cercano a -1: se predice lo contrario

| prediction | actual value | | | |
|------------|--------------|---------|---------|-----------|
| | + | - | Sum | |
| | Pos | a | b | T_{pos} |
| | Neg | c | d | T_{neg} |
| | Sum | T_{+} | T_{-} | T |

note:

+/- are actual values

pos/neg are predictions

Averiguar

Kappa coefficient

Accuracy ignora diferentes costos de clasificaciones incorrectas

Accuracy ignora la posibilidad de tener diferentes costos por clasificación

- A veces clasificar incorrectamente “+” cuesta más/menos que clasificar incorrectamente “-”
- No tratar un paciente enfermo vs tratar a uno sano
- Negar crédito a un usuario que si paga vs darle crédito a uno que no

Se necesita investigar la probabilidad de tener diferentes tipos de errores

Una posible solución es distinguir “predictive accuracy” para las diferentes clases

Acc: probabilidad de que alguna instancia se clasifique correctamente

- TP: true positive rate, probabilidad de que una instancia positiva se clasifique correctamente
- TN: true negative rate, probabilidad de que una instancia negativa se clasifique correctamente

También se define

- $FP=1-TN$: false positive rate, probabilidad que una instancia negativa se clasifique como positiva
- $FN=1-TP$, false negative rate

Una posible solución es distinguir “predictive accuracy” para las diferentes clases

Se consideran los costos C_{FP} y C_{FN}

- Costos de falsos positivos y falsos negativos

El costo esperado de una predicción

- $C = C_{FP} P(pos|-) P(-) + C_{FN} P(neg|+) P(+)$
- Donde $C = C_{FP} FP T_- / T + C_{FN} FN T_+ / T$

Nótese

- Acc es el promedio ponderado de TP y TN: $Acc = TP T_+ / T + TN T_- / T$
- C no se calcula de Acc

Accuracy es sensible al cambio de la distribución de clases

Si la distribución de clases en el test set difiere de la del entrenamiento, el Acc también será diferente

Suponer que un clasificador tiene $TP = 0.8$, $TN = 0.6$

Probado en un test set con $T+/T = 0.5$, $T-/T = 0.5$ ($Acc=0.7$)

Empleado con nuevos datos con $T+/T = 0.3$, $T-/T = 0.7$ ($Acc=0.66$)

Receiver operating characteristic (ROC) permite analizar como un clasificador se comporta

Permite

- Cuan bien se comporta un clasificador dados los costos de clasificaciones incorrectas y distribución de clases
- En que ambientes un clasificador es mejor que otro

Provost & Fawcett (2001), Robust classification under imprecise environments, Machine Learning 42(3):203-231

Un **diagrama ROC**

muestra TP vs FP

Dada una matriz de confusión

$$TP = a/(a+c) = a/T_+$$

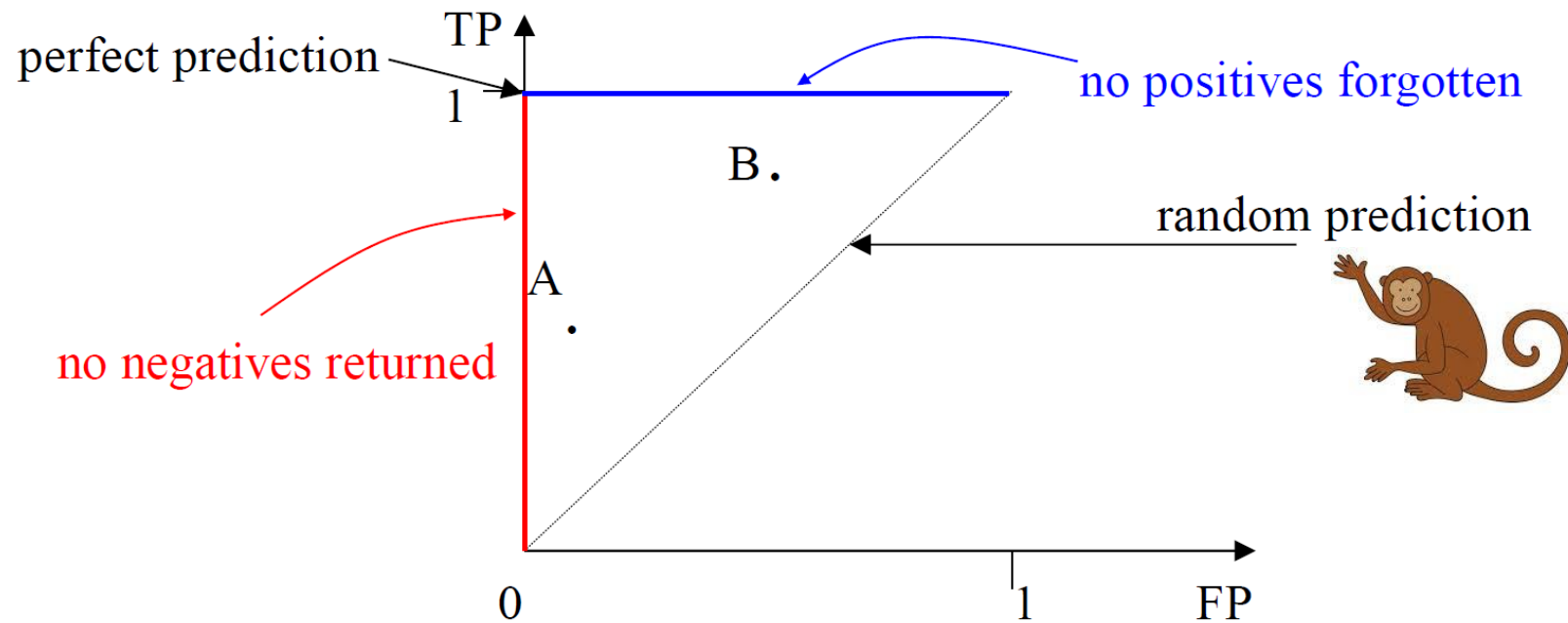
$$FP = b/(b+d) = b/T_-$$

| prediction | actual value | | | |
|------------|--------------|-------|-------|------------------|
| | | + | - | Sum |
| | Pos | a | b | T_{pos} |
| | Neg | c | d | T_{neg} |
| | Sum | T_+ | T_- | T |

Un diagrama ROC debe ser interpretado

1 clasificador = 1 punto en el diagrama

Mientras más cerca a la esquina superior izquierda, mejor



Algunos clasificadores proveen una medida de certeza en sus predicciones

Algunas predicciones tiene más certeza que otras -> rango más alto

Ejemplo: Redes Neuronales

- Criterio: predicción $< 0.5 \rightarrow neg$, $\geq 0.5 \rightarrow pos$
- 0.9 tiene más certeza que 0.51
- El threshold puede modificarse, por defecto 0.5. Threshold más alto -> menos positivos (TP y FP decrecen)

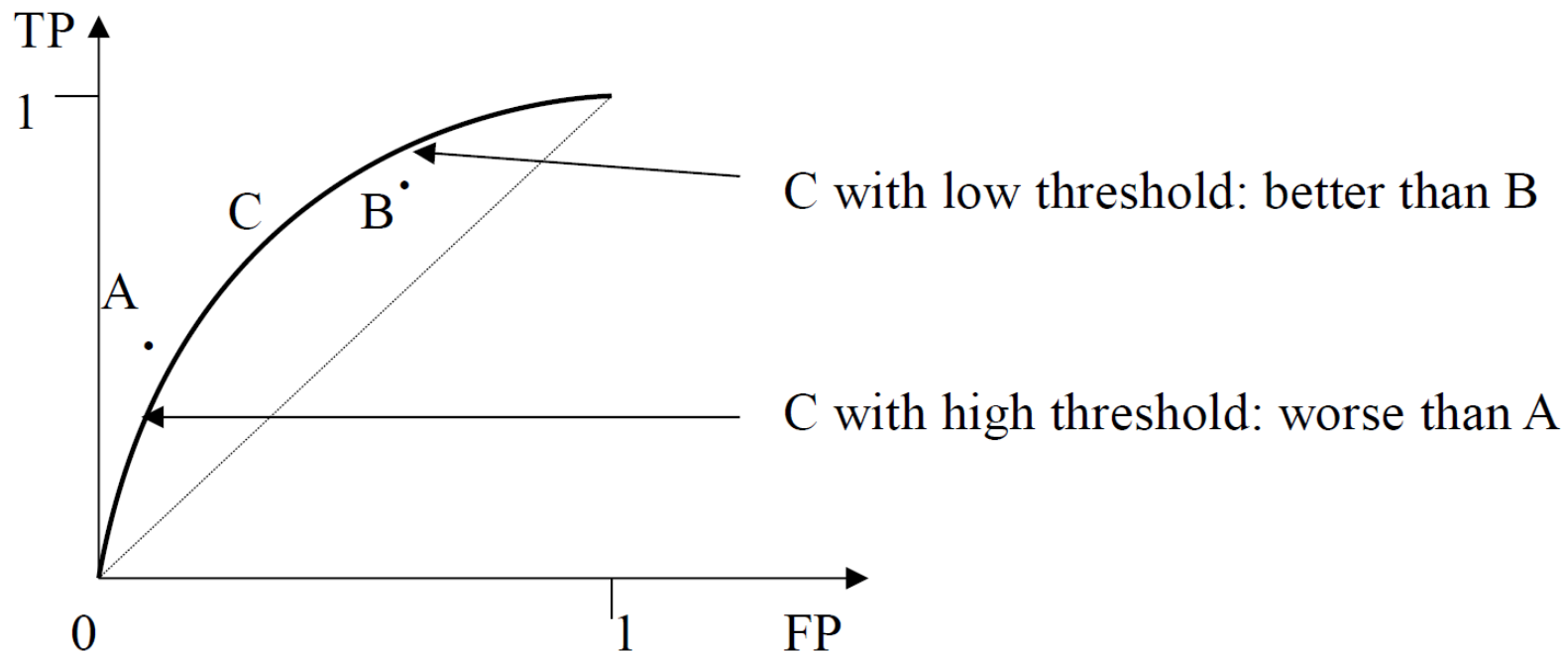
Ejemplo: Arboles de Decisión

- Se usa la pureza de las hojas para determinar la certeza
- Una hoja con 90% de positivos tiene más certeza que una con 80% de positivos

Este tipo de clasificadores posibilitan la creación de una curva ROC

Con un valor específico de threshold: 1 punto en la curva

La curva se dibuja en función del threshold especificado $C(th)$



Un diagrama ROC también puede mostrar los costos de clasificaciones incorrectas

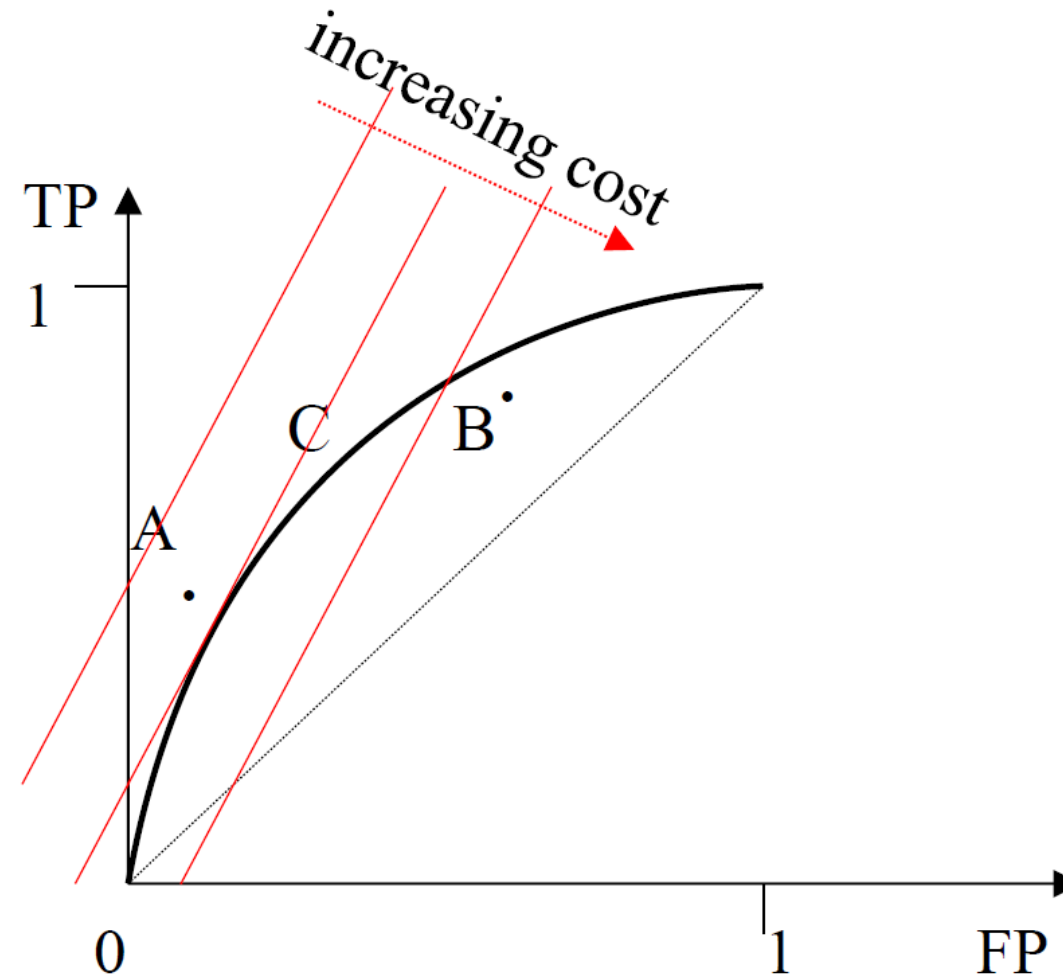
Dados los costos de clasificaciones incorrectas

- C_{FP} : costo de un falso positivo
- C_{FN} : costo de un falso negativo (+ no detectado)

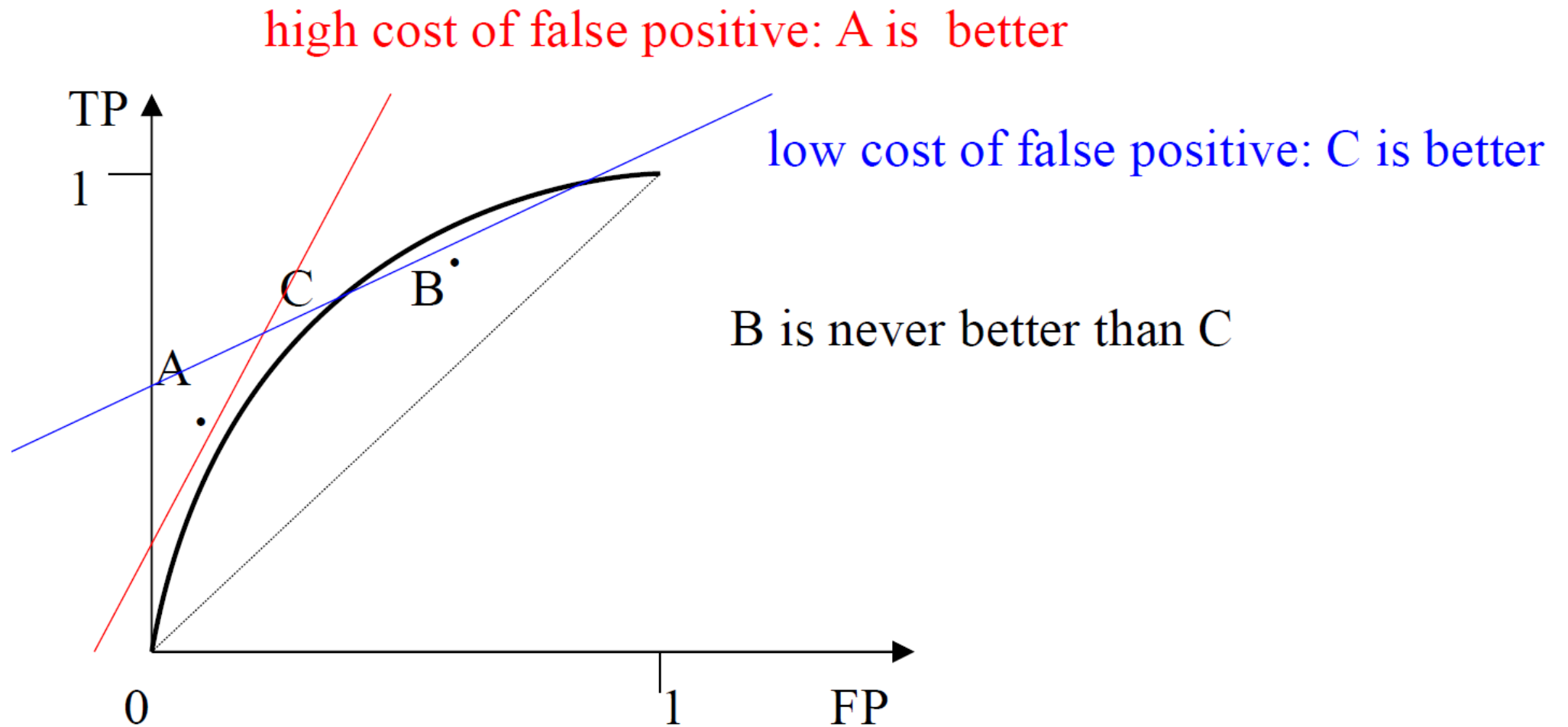
El costo promedio es

- $c = C_{FP} \times FP \times \frac{T_-}{T} + C_{FN} \times (1 - TP) \times \frac{T_+}{T}$
- Líneas de igual costo pueden ser dibujadas en el diagrama ROC
- La pendiente de la línea: $\frac{C_{FP} \times \frac{T_-}{T}}{C_{FN} \times \frac{T_+}{T}}$

Un diagrama ROC también puede mostrar los costos de clasificaciones incorrectas



Un diagrama ROC también puede mostrar
los costos de clasificaciones incorrectas



Diferentes modelos pueden ser buenos en diferentes ambientes

Por ejemplo:

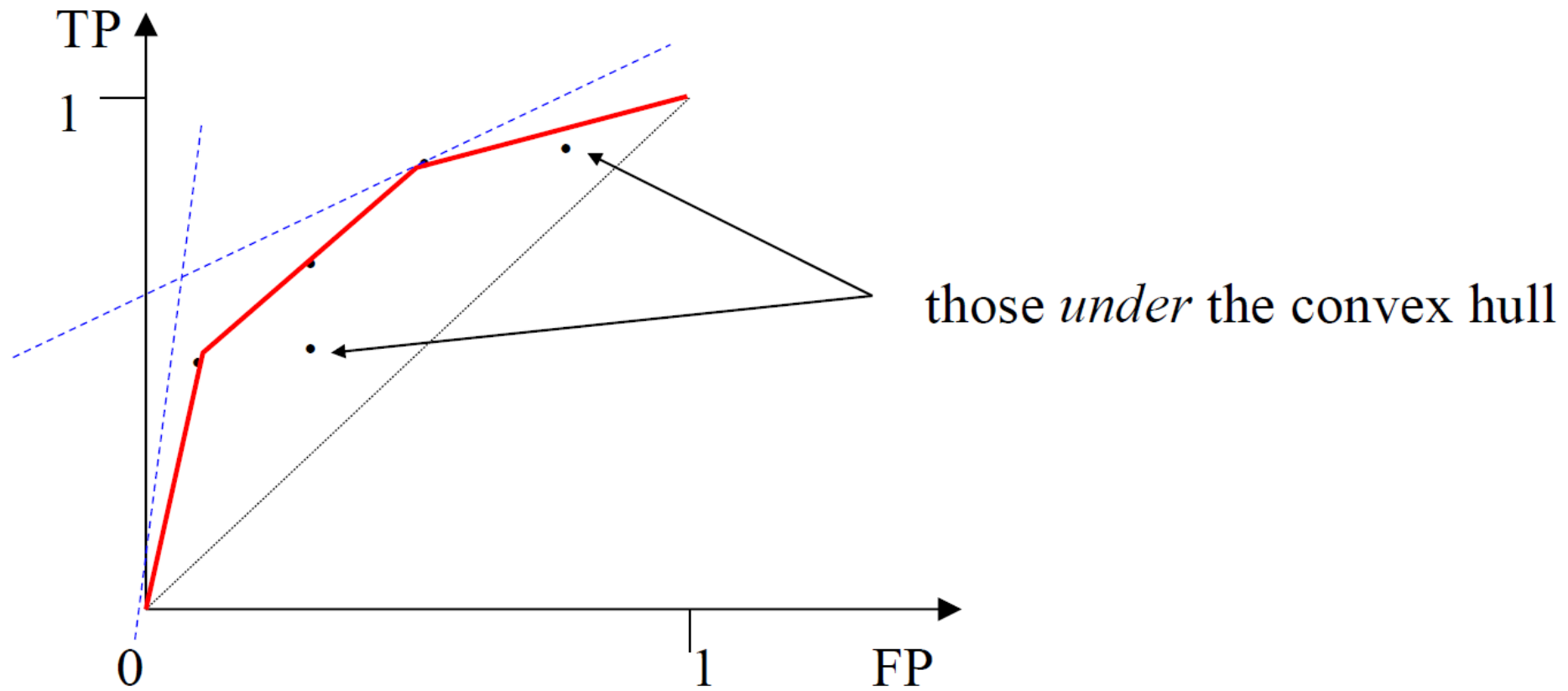
- Diferentes costos de clasificación incorrecta
- Diferentes distribuciones de clase

Dado un conjunto de modelos, un análisis ROC permite:

- Decidir en que casos un modelo es óptimo
- Remover modelos que nunca serán óptimos

Modelos que pueden ser óptimos están siempre sobre el “convex hull” del conjunto de puntos

Ejemplo: ¿Qué clasificadores
nunca son óptimos?



Content



Comparar modelos

Técnicas para comparar calidad de soluciones

Métricas de Clasificación

Técnicas clásicas e interpretación

Métricas de Regresión

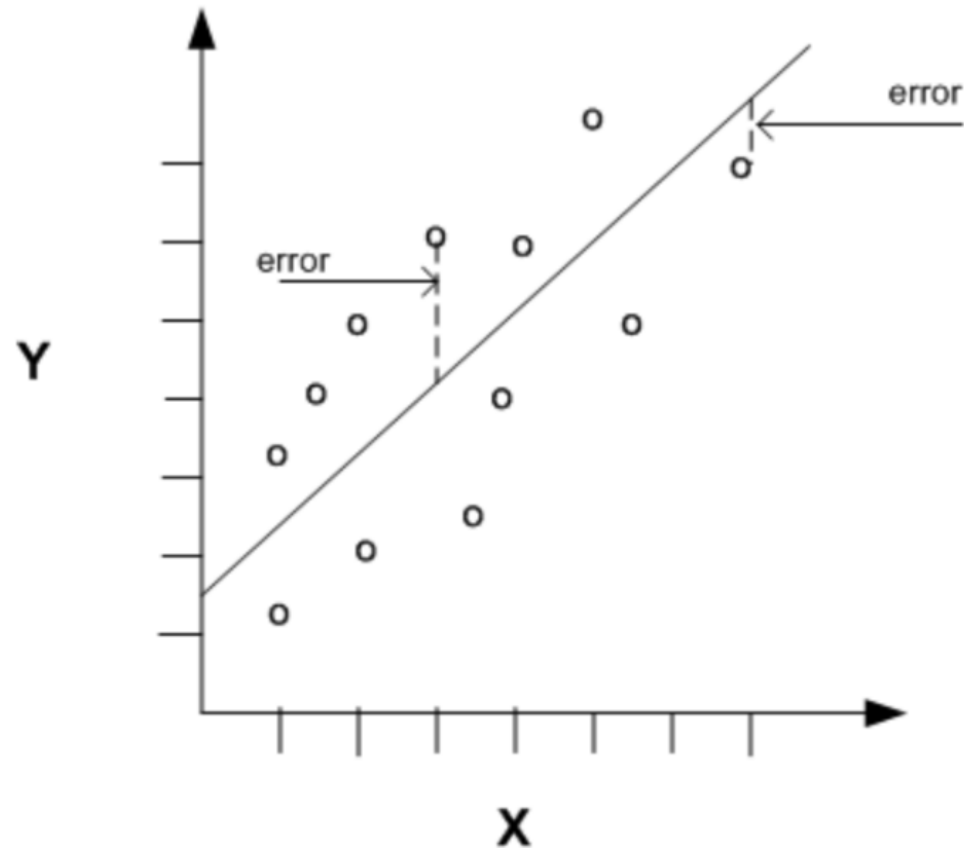
Tipos de métricas e interpretación

Los modelos de regresión también deben ser evaluados

Posibles medidas:

- Suma de errores al cuadrado SSE (medida absoluta)
- Error relativo RE : mide la mejora sobre un modelo trivial
$$RE = \frac{SSE_{hypothesis}}{SSE_{trivialHypothesis}}$$
- Ejemplos de hipótesis trivial: predecir siempre la media, modelo persistente, etc.
- Spearman correlation r : mide cuan bien las predicciones y valores verdaderos se correlacionan (menos sensible a errores verdaderos)

Mean Absolute Error (MAE) es el promedio de los errores producidos por el modelo de regresión



$$MAE = \text{mean}(|e_i|)$$

No se suele utilizar esta medida para entrenamiento ya que no tiene una forma cuadrática

La unidad de error es la misma que la del dataset

Mean squared error (MSE) resuelve el problema de la diferenciabilidad de MAE

$$MSE = \frac{1}{n} \sum_{i=1}^t e_i^2$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^t e_i^2}$$

Al estar elevadas al cuadrado, estas medidas penalizan más a los errores más altos

RMSE penaliza los errores más grandes

Ejemplo

CASE 1: Evenly distributed errors

| ID | Error | Error | Error^2 |
|----|-------|-------|---------|
| 1 | 2 | 2 | 4 |
| 2 | 2 | 2 | 4 |
| 3 | 2 | 2 | 4 |
| 4 | 2 | 2 | 4 |
| 5 | 2 | 2 | 4 |
| 6 | 2 | 2 | 4 |
| 7 | 2 | 2 | 4 |
| 8 | 2 | 2 | 4 |
| 9 | 2 | 2 | 4 |
| 10 | 2 | 2 | 4 |

| MAE | RMSE |
|-------|-------|
| 2.000 | 2.000 |

CASE 2: Small variance in errors

| ID | Error | Error | Error^2 |
|----|-------|-------|---------|
| 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 1 |
| 3 | 1 | 1 | 1 |
| 4 | 1 | 1 | 1 |
| 5 | 1 | 1 | 1 |
| 6 | 3 | 3 | 9 |
| 7 | 3 | 3 | 9 |
| 8 | 3 | 3 | 9 |
| 9 | 3 | 3 | 9 |
| 10 | 3 | 3 | 9 |

| MAE | RMSE |
|-------|-------|
| 2.000 | 2.236 |

CASE 3: Large error outlier

| ID | Error | Error | Error^2 |
|----|-------|-------|---------|
| 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 |
| 10 | 20 | 20 | 400 |

| MAE | RMSE |
|-------|-------|
| 2.000 | 6.325 |

Mean Absolute Percentage Error (MAPE) es una de las medidas más ampliamente utilizadas para comparar modelos

$$MAPE = 100 \times \text{mean}\left(\frac{|e_i|}{y_i}\right)$$

- Su escala va de 0-100
- Debido a esta escala, se utiliza ampliamente para comparar modelos
- Tiende al infinito si y_i es muy pequeña o es 0

Mean Absolute Scaled Error (MASE) es la proporción entre MAE y un naive forecast calculado con los datos de entrenamiento

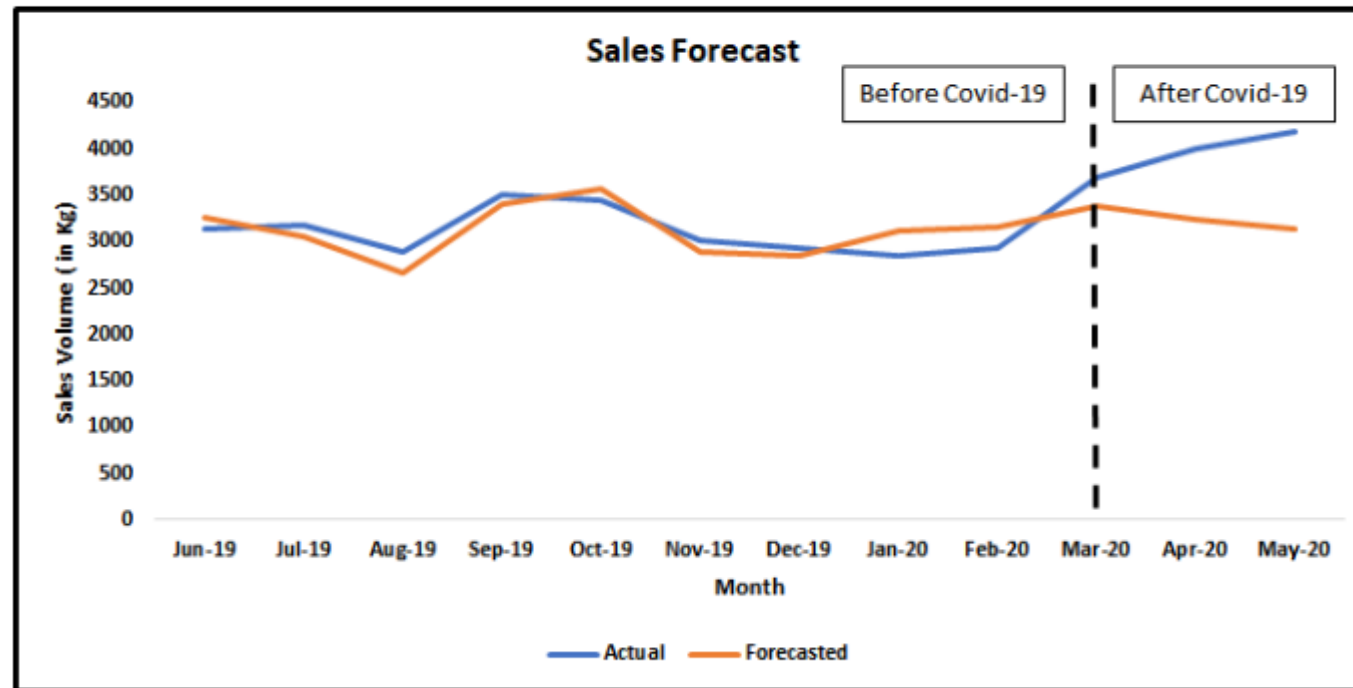
$$MASE = \frac{MAE}{Q}$$

$$Q = \frac{1}{N-1} \sum_{j=2}^N |y_j - y_{j-1}|$$

- MAE es calculado sobre el test set y Q sobre el training set
- Si $MASE < 1$, el modelo es mejor que un forecast producido por un naive model

Ejemplo: ¿Qué métrica utilizar?

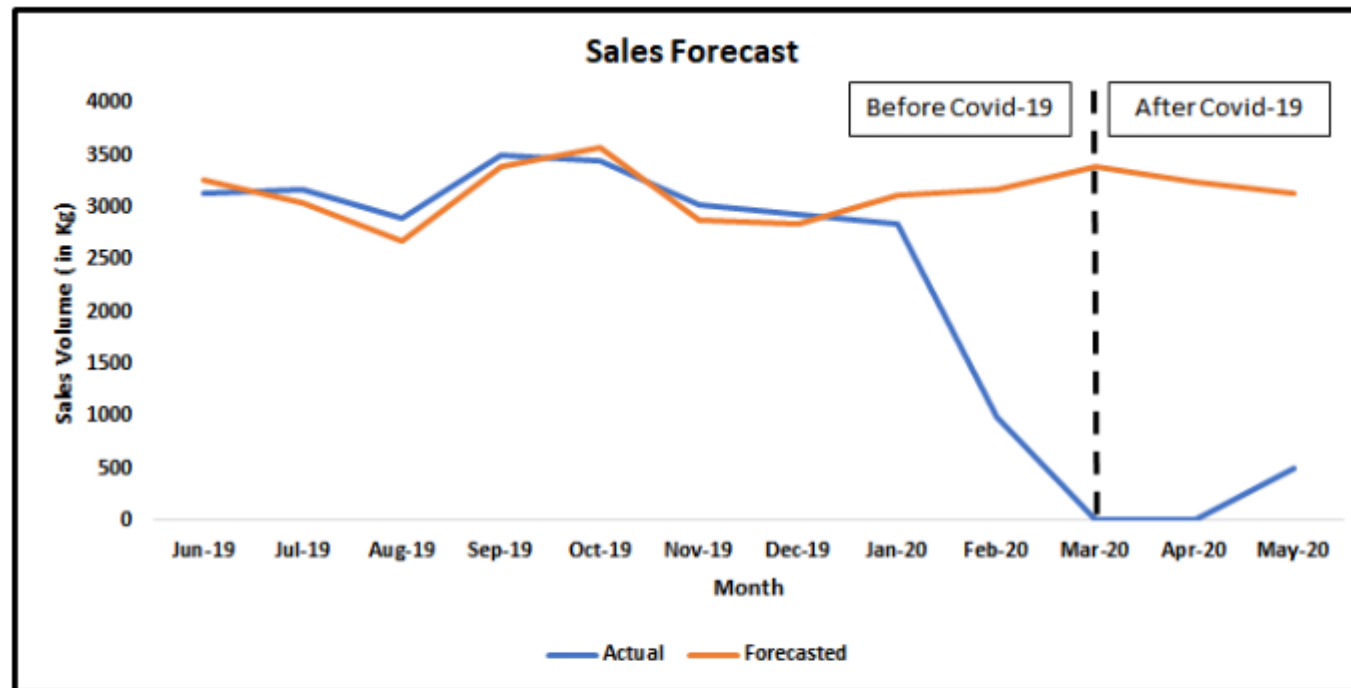
Scenario 1: Forecast varying drastically from actuals



Al ser independiente de la escala, MAPE podría ser utilizado para este escenario

| | MAE | RMSE | MAPE | MASE |
|-----------|--------|--------|--------|------|
| Test data | 703.67 | 768.39 | 17.38% | 3.39 |

Ejemplo: ¿Qué métrica utilizar?

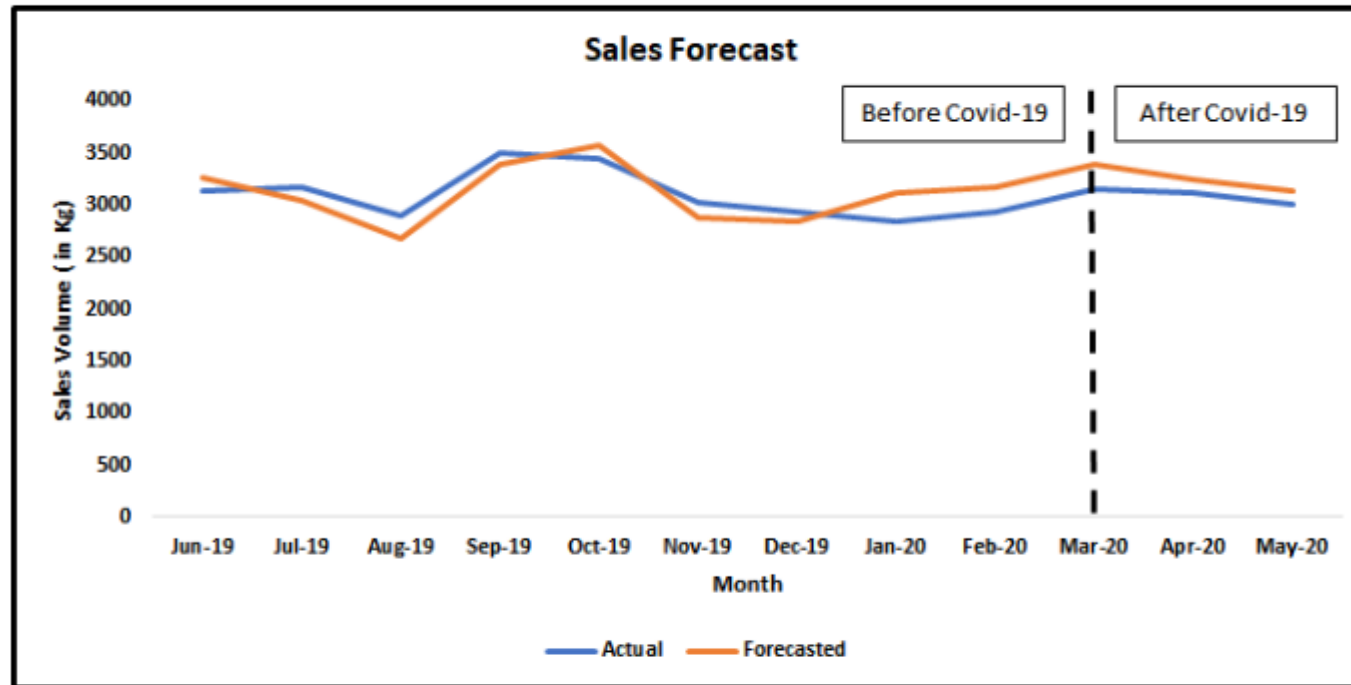


Para este caso MAE o RMSE son medidas que pueden ser usadas para comparar modelos

| | MAE | RMSE | MAPE | MASE |
|-----------|---------|---------|---------|------|
| Test data | 3096.33 | 3113.44 | #DIV/0! | 7.21 |

Ejemplo: ¿Qué métrica utilizar?

Scenario 3: The ideal scenario (Back to normal scenario)



En condiciones normales, cualquier métrica puede ser utilizada

| | MAE | RMSE | MAPE | MASE |
|-----------|--------|--------|-------|------|
| Test data | 164.00 | 172.58 | 5.27% | 0.79 |

Coeficiente de determinación R^2 , es otra métrica comúnmente utilizada para analizar regresiones

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$

Mide la cantidad de varianza de predicción que es explicada por el modelo

En otras palabras, mide cuan bien las predicciones se asemejan a los valores verdaderos

Coeficiente de determinación R^2 , es otra métrica comúnmente utilizada para analizar regresiones

$$R^2 = 1 - \frac{\sum_{i=1}^t e_i^2}{\sum_{i=1}^t (y_i - \bar{y})^2}$$

Mide si el modelo implementado tiene mayor error que un modelo lineal simple

$$R^2 = 1 - \frac{\text{Error del modelo}}{\text{Error modelo promedio}}$$

Donde, modelo promedio es uno que solo predice el valor promedio del dataset



UNIVERSIDAD DE CUENCA
desde 1867

Evaluación de la Hipótesis

Andres Auquilla

2024