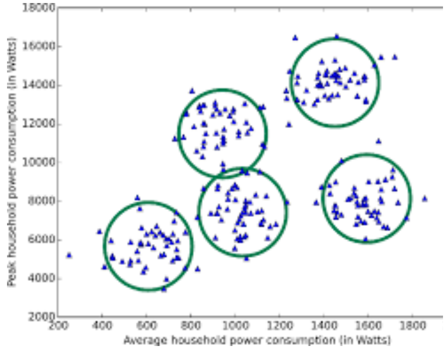
APRENDIZAJE SUPERVISADO

Anibal Sosa, PhD

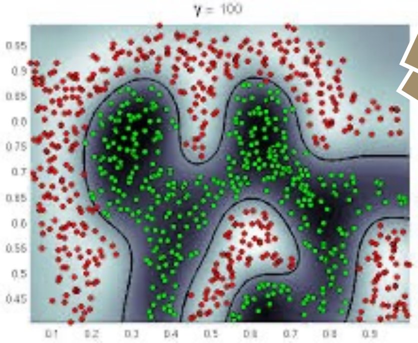
UNIVERSIDAD ICESI

Clase anterior

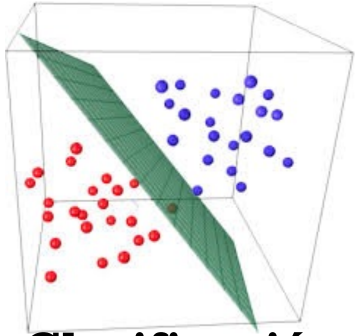
Aprendizaje automático



Aprendizaje no supervisado

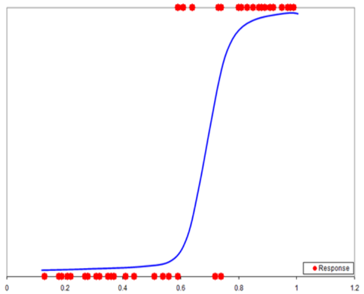


Aprendizaje supervisado

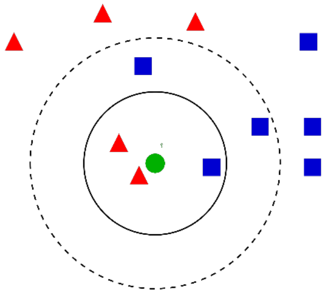


Clasificación

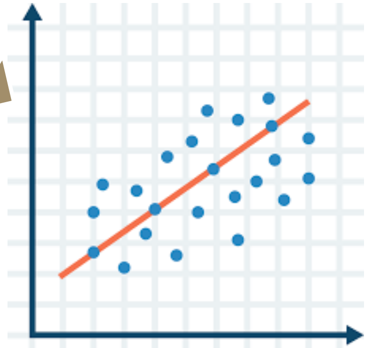
Métricas de Evaluación de la clasificación



Regresión logística



KNN



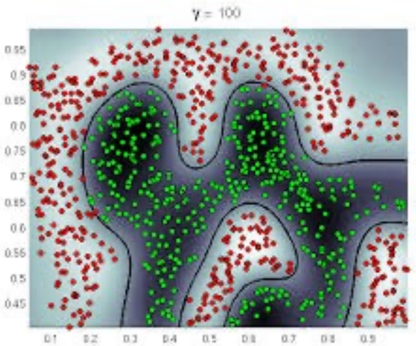
Regresión

Métricas de Evaluación de la regresión

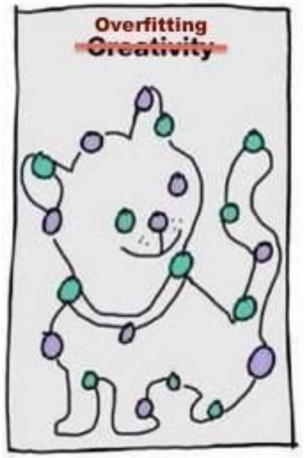
UNIVERSIDAD ICESI

AGENDA

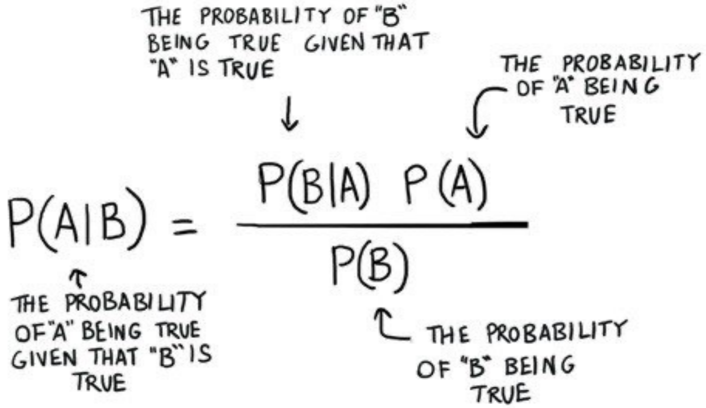
Protocolos



Aprendizaje supervisado



Sobre aprendizaje (Overfitting)



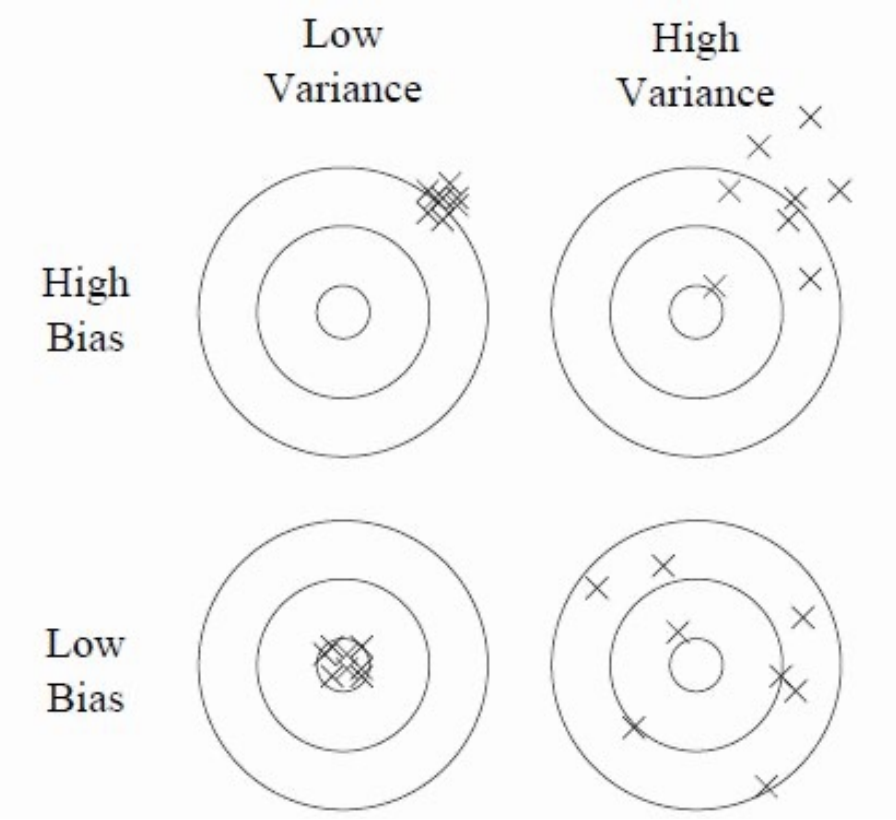
Naïve Bayes

OVERFITTING Y PROTOCOLOS DE EVALUACIÓN

UNIVERSIDAD
ICESI

SESGO/ VARIANZA

- **Sesgo** (bias): que tan lejos está el modelo de la verdad
- **Varianza**: Qué tanto varían los datos de la predicción para una misma instancia

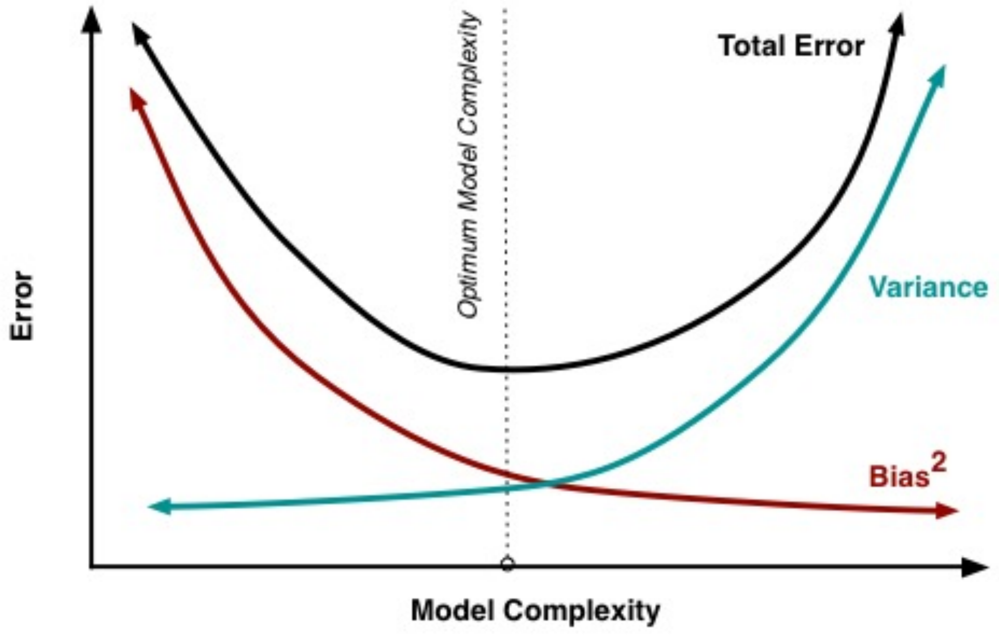


Domingo, 2012

$$Err(x) = \underbrace{\left(\overset{\text{Promedio del modelo}}{E\left[\hat{f}(x)\right]} - \overset{\text{Verdad}}{f(x)}\right)^2}_{\text{Sesgo}^2} + \underbrace{E\left[\overset{\text{cada resultado del modelo}}{\hat{f}(x)} - \overset{\text{Promedio del modelo}}{E[\hat{f}(x)]}\right]^2}_{\text{Varianza}} + \underbrace{\sigma_e^2}_{\text{Error irreducible}}$$

SESGO/ VARIANZA

- Ambos son fuente de error
- Se debe determinar un **compromiso** entre ambos tipos de error
- Parámetros de los modelos controlan la complejidad



Bias / Variance

SOBRE APRENDIZAJE (OVERFITTING)

¿Cómo le enseño a un niño que es una pelota?

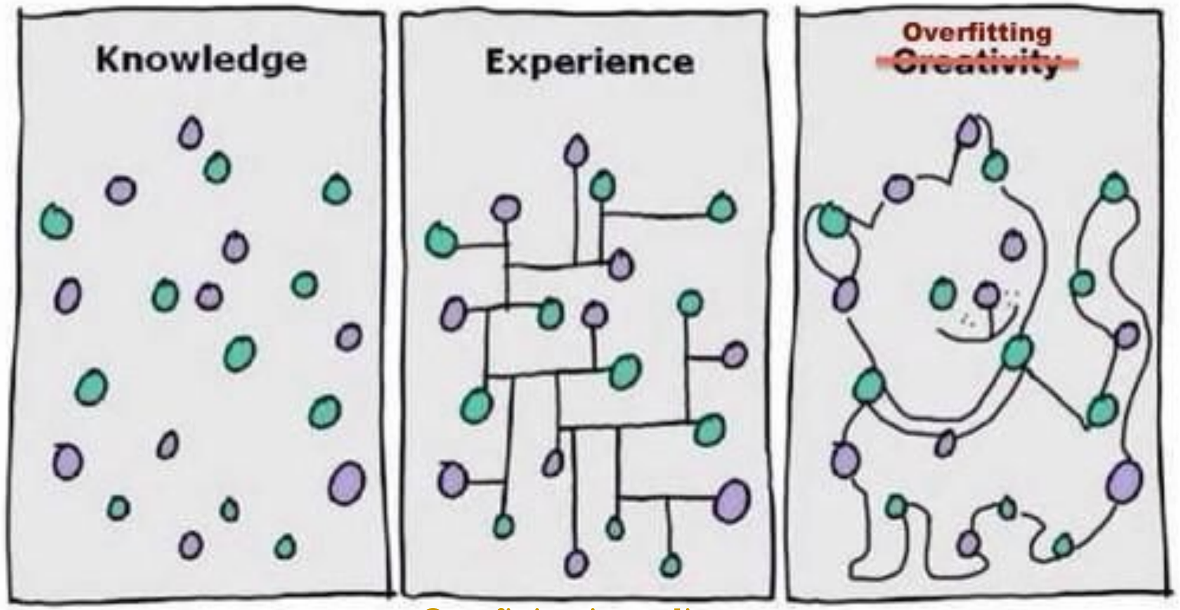
Set de entrenamiento

¿Qué patrones distinguen las pelotas de los demás juguetes?

¿Es ésta una pelota?

¿Cómo caracterizo una situación de un modelo con overfitting?

SOBRE APRENDIZAJE (OVERFITTING)

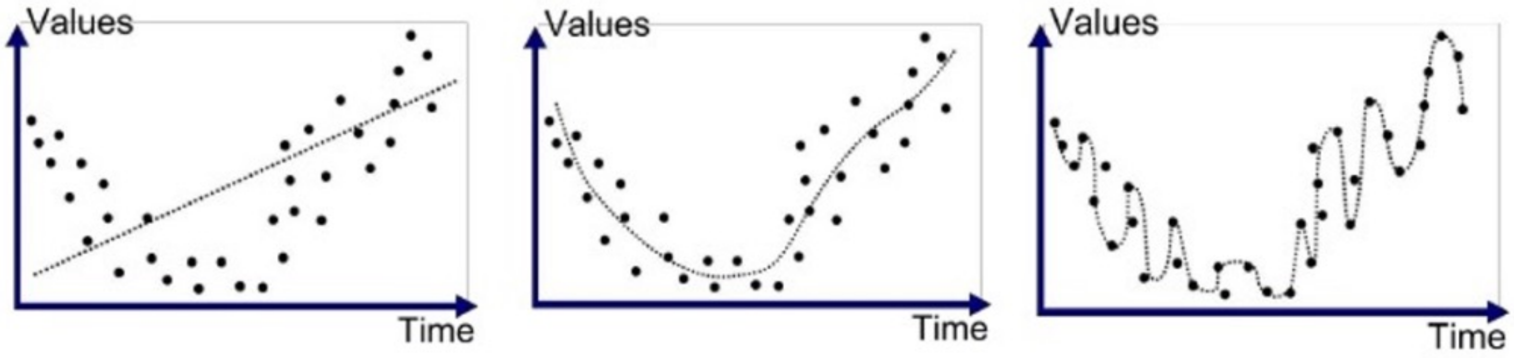


Overfitting in trading

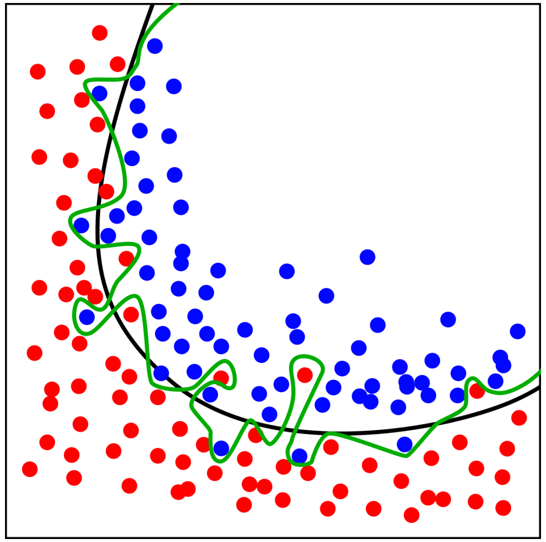
- **Sobre aprendizaje**: Los modelos aprenden a describir los errores aleatorios o el "ruido" del conjunto de entrenamiento.
- Ocurre cuando un modelo se vuelve excesivamente **complejo**

SOBRE APRENDIZAJE (OVERFITTING)

Regresión



Clasificación



Overfitting

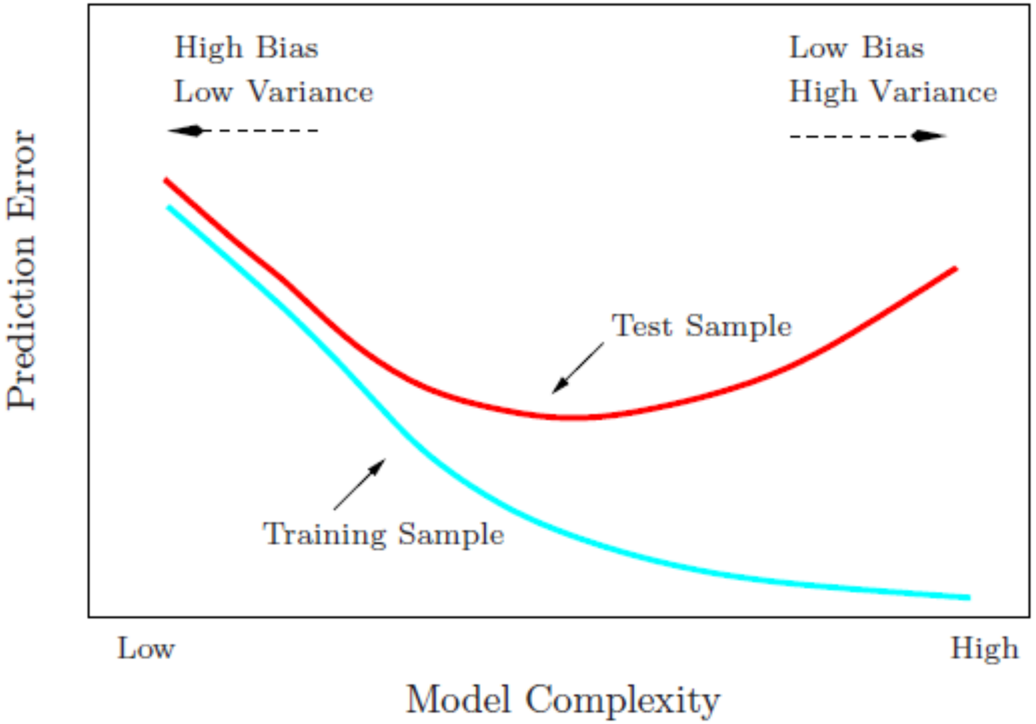
¿Cómo es el sesgo y la varianza de estos modelos?

- La **complejidad** de un modelo debe ajustarse de tal manera que permita la **generalización**, al utilizarse con datos que no haya conocido durante el proceso de entrenamiento
- Principio de **parsimonia** (Occam's Razor): la mejor explicación es la más simple → preferir los modelos más sencillos con menos suposiciones

UNIVERSIDAD ICESI

SOBRE APRENDIZAJE (OVERFITTING)

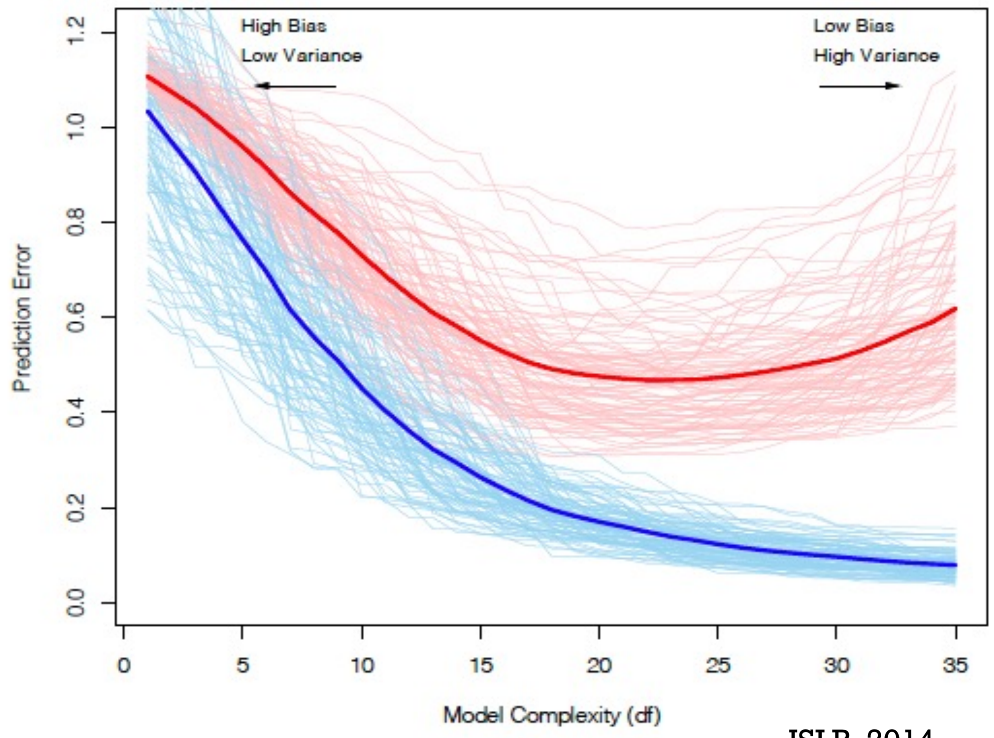
- Los modelos tienden a ajustarse al conjunto de datos usado para su aprendizaje → el **error de entrenamiento** es un mal estimador
- Queremos encontrar la complejidad del modelo que nos permita minimizar el **error de test**



Overfitting

SOBRE APRENDIZAJE (OVERFITTING)

- Los modelos tienden a ajustarse al conjunto de datos de entrenamiento → el **error de entrenamiento** es un mal estimador
- Queremos encontrar la complejidad del modelo que nos permita minimizar el **error de prueba**

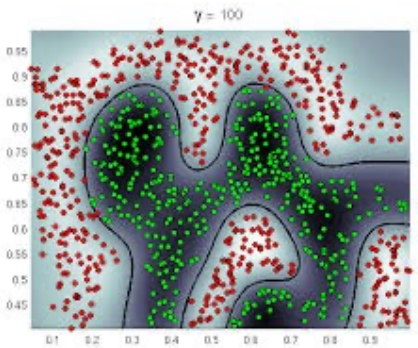


ISLR, 2014

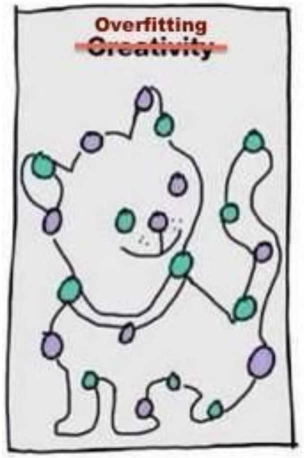
Double Descent

AGENDA

Protocolos



Aprendizaje supervisado

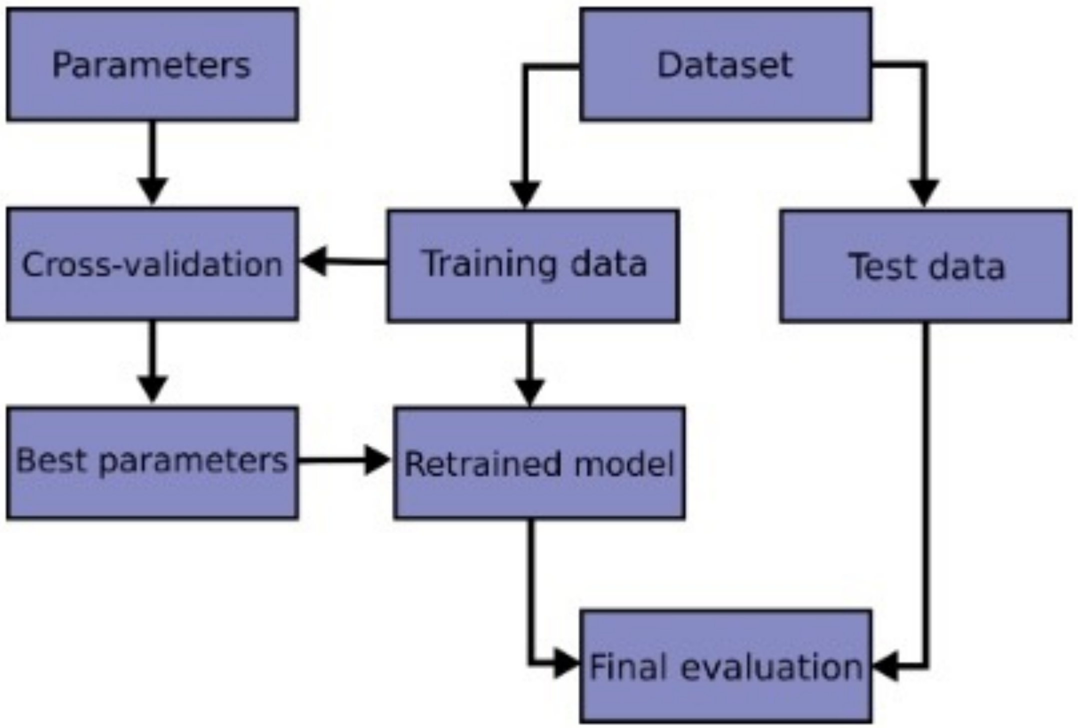


Sobre aprendizaje (Overfitting)

PROTOCOLOS DE EVALUACIÓN

- Aplican para aprendizaje supervisado en general.
- Permiten:
  - Definir los mejores valores de los hiper parámetros de los modelos (**flexibilidad)**.
  - Evaluar la capacidad de **generalización** del modelo.
  - Comparar el **error de entrenamiento** y el **error de test.**
  - Evitar el sesgo causado por la **subestimación del error** al evaluar con el mismo set de entrenamiento.
  - Establecer un compromiso entre sesgo y varianza, para reducir el **sobre aprendizaje** y encontrar un modelo con buenas **capacidades predictivas.**

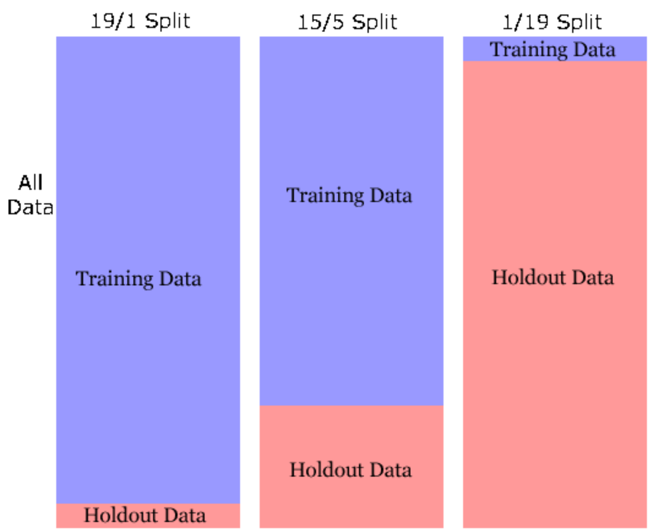
PROTOCOLOS DE EVALUACIÓN



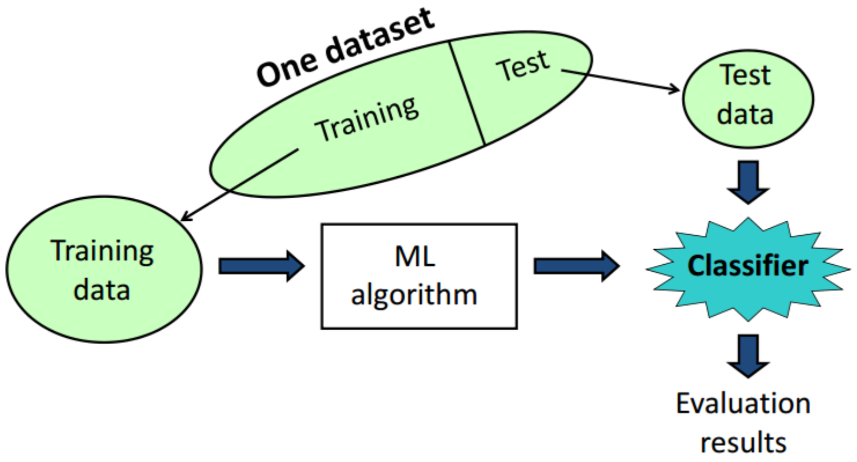
Cross - validation

PROTOCOLOS DE EVALUACIÓN

- **Holdout**: particionar el conjunto de datos en dos:
  - **Conjunto de entrenamiento (*train*)**: con el que aprende el algoritmo de clasificación
  - **Conjunto de validación (*test*)**: separado al comienzo del proceso y no considerado en el aprendizaje
  - **Aleatoriedad** del particionamiento
  - **Compromiso**: entre más datos mejor el aprendizaje, y la evaluación
- **Repeated holdout**: repetir el procedimiento y agregar las métricas de evaluación



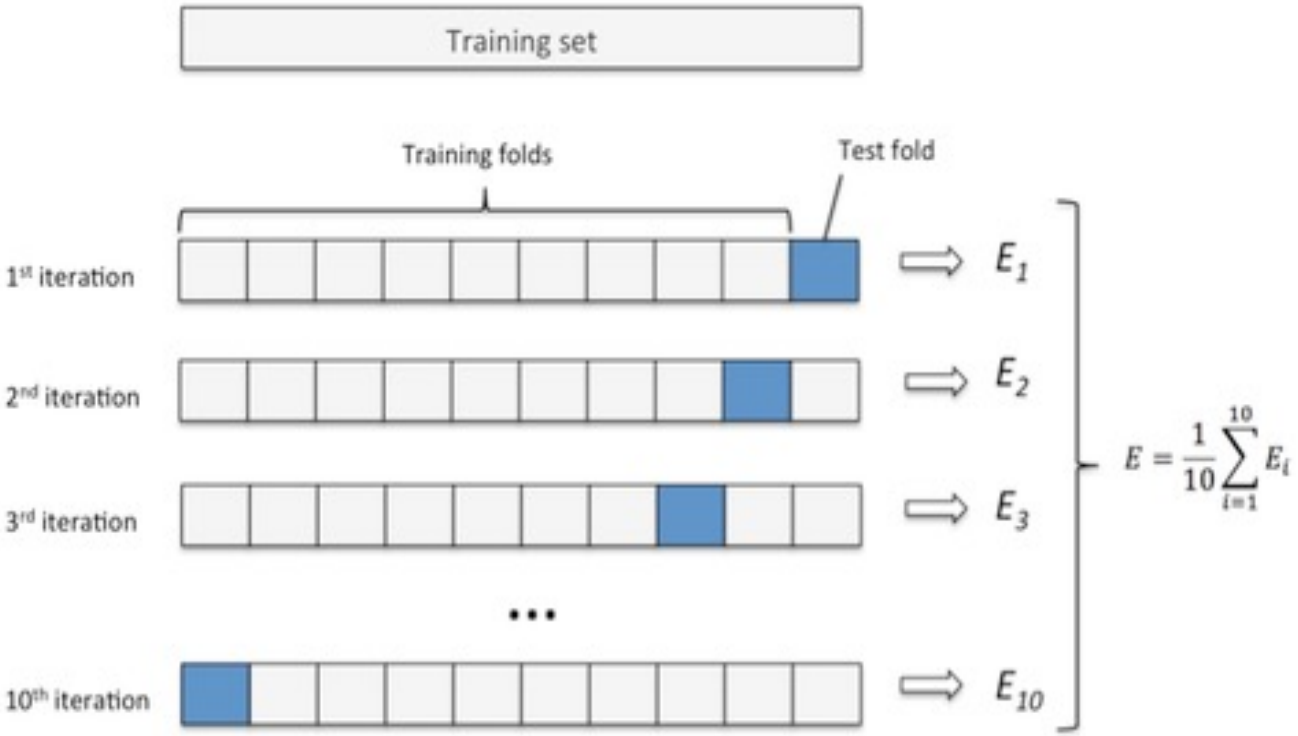
Holdout



Ian Witten, Weka MOOC

PROTOCOLOS DE EVALUACIÓN

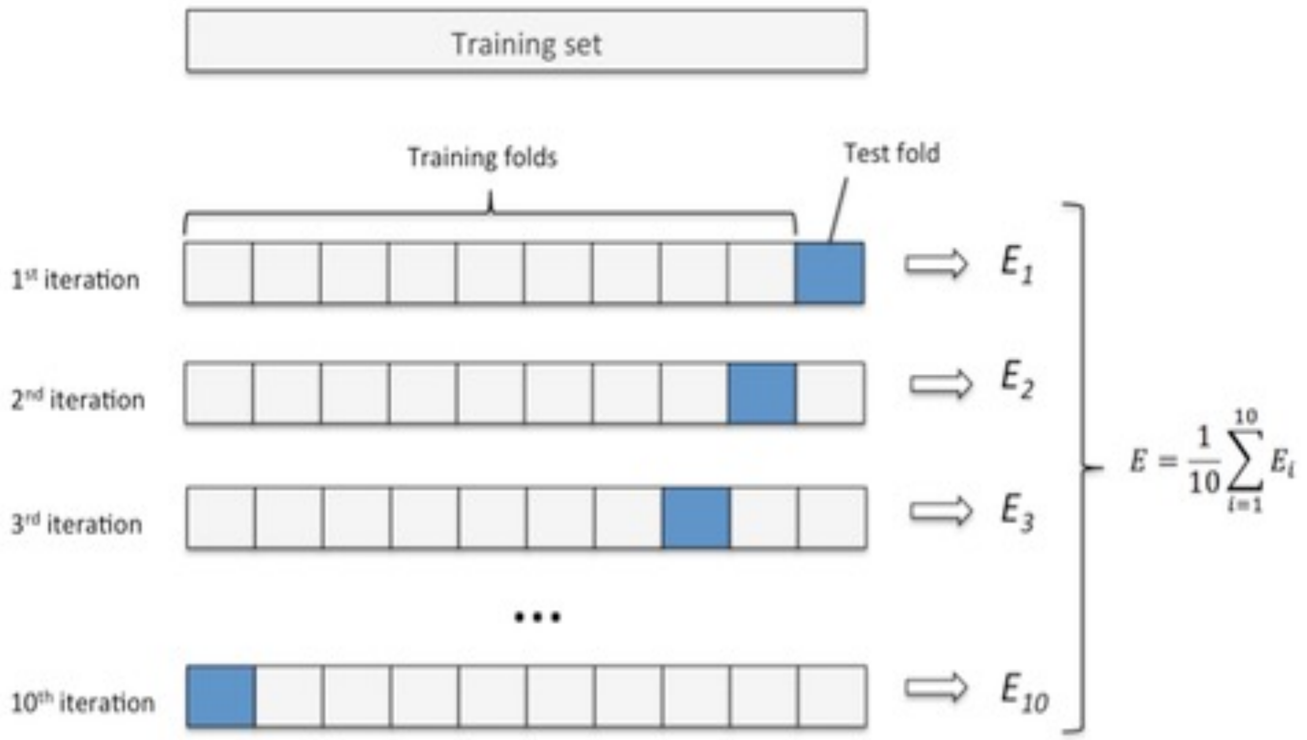
- **K-fold cross-validation**:
  - Particionar el conjunto de datos en K conjuntos disjuntos del mismo tamaño
  - K-1 partes se usan para entrenamiento, una parte se usa para el test
  - Se repite el proceso K veces
  - Se agregan las métricas de evaluación



Sebastian Raschka, 2015

PROTOCOLOS DE EVALUACIÓN

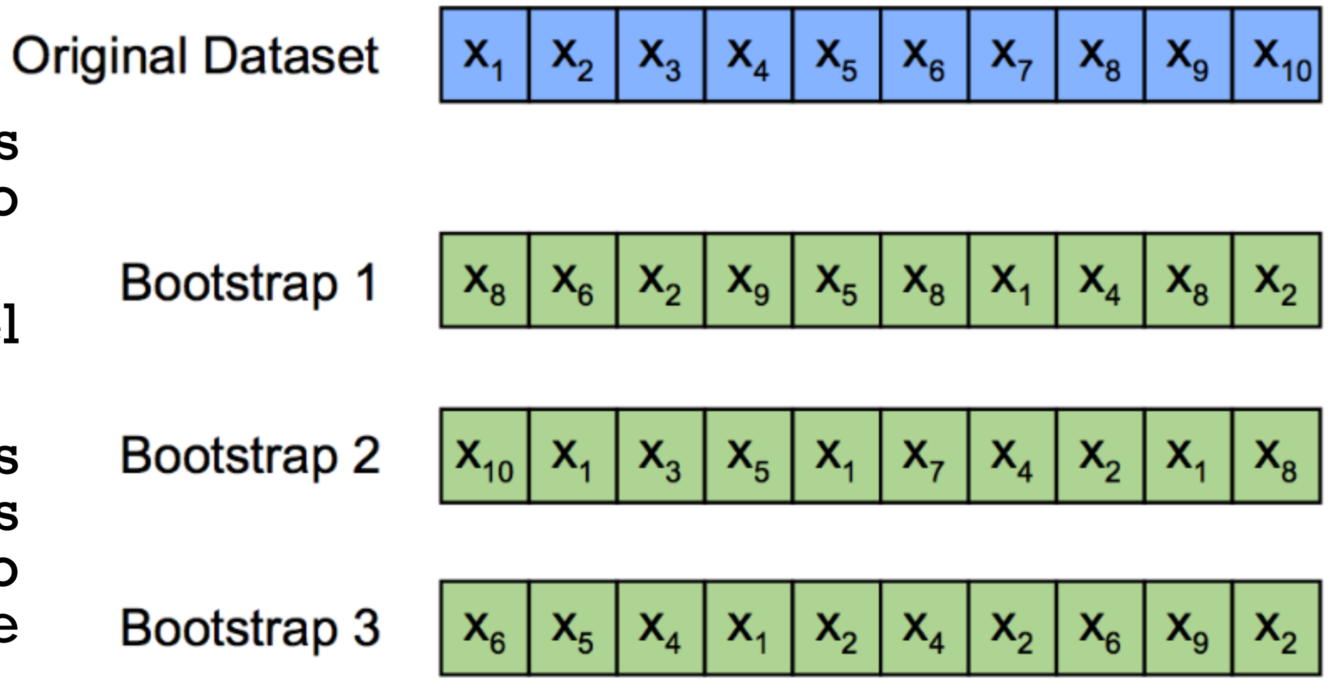
- **K-fold cross-validation, Escogencia del K**:
  - Permite balancear entre sesgo y varianza
  - **LOOCV** (Leave One Out Cross-Validation): conjuntos unitarios
  - Se estima que los mejores resultados se obtienen con un valor de K entre 5 y 10



Sebastian Raschka, 2015

PROTOCOLOS DE EVALUACIÓN

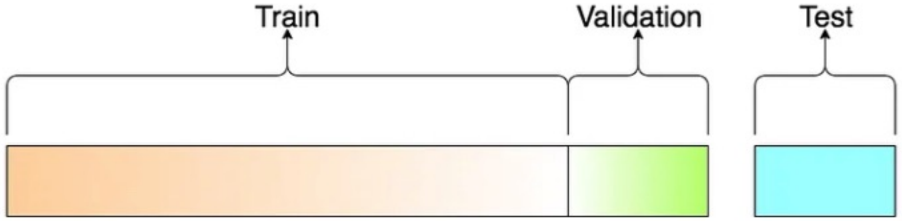
- **Bootstrapping**:
  - Consideración de varios conjuntos de entrenamiento/test utilizando muestreo con remplazo
  - Por lo general, los muestreos son del mismo tamaño del conjunto original
  - Muy buen estimador de los parámetros, pero no de las métricas de calidad de los modelos (sesgo causado por el promedio de observaciones distintas (0.632*N))



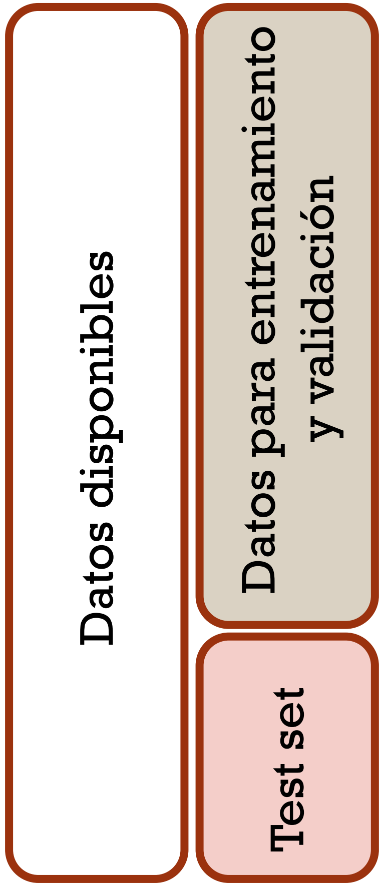
Sebastian Raschka, 2015

PROTOCOLOS DE EVALUACIÓN

- Set de validación vs set de test:
  - Separación de un set de datos de **test** para evaluación final del modelo escogido
  - **Overfitting** si se calibran los modelos con el mismo set de test



Toward data science



- Calibración de pretratamientos (normalización, imputación)
- Calibración de los parámetros de los modelos (KFCV, holdout)
- Comparación de los resultados de diferentes modelos

- Evaluar la capacidad de generalización

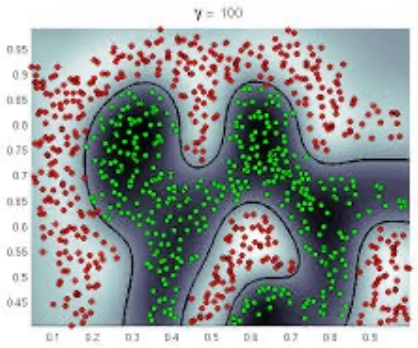
EJEMPLO DE KNN

- 04-KNN-Protocolos-Ejemplo
- **Desarrollo de k-nn sobre el dataset de iris utilizando diversos protocolos de evaluación**

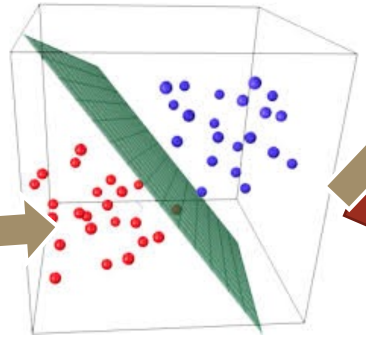
TALLER DE EVALUACIÓN DE UN MODELO DE CLASIFICACIÓN (KNN)

- DATASET (04-02-churn.csv): base de datos de 20000 clientes que han cancelado (churn) o no los servicios de una compañía. La idea es poder predecir en un futuro quiénes son los clientes más propensos a hacer churn, para poder desarrollar campañas que lo prevengan.
- Descargue el archivo 04-KNN-CHURN y ejecútelo, vamos a ir revisando por partes.

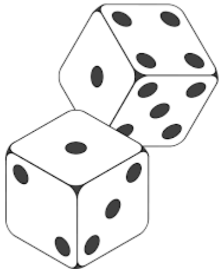
UNIVERSIDAD ICESI



Aprendizaje supervisado



Clasificación



Probabilidades

THE PROBABILITY OF "B" BEING TRUE GIVEN THAT "A" IS TRUE

THE PROBABILITY OF "A" BEING TRUE

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

THE PROBABILITY OF "A" BEING TRUE GIVEN THAT "B" IS TRUE

THE PROBABILITY OF "B" BEING TRUE

Naïve Bayes

PRINCIPIOS BASICOS DE PROBABILIDAD

Probabilidad Marginal:

$$P(X)$$

Regla del producto: *(Probabilidad Conjunta)*

$P(X \cap Y) = P(X) * P(Y)$ (si X y Y son eventos independientes)

Ley de probabilidad total:

$P(X) = \sum_n P(X|Y_n)\, P(Y_n)$ (si X y $Y_n$son eventos independientes)

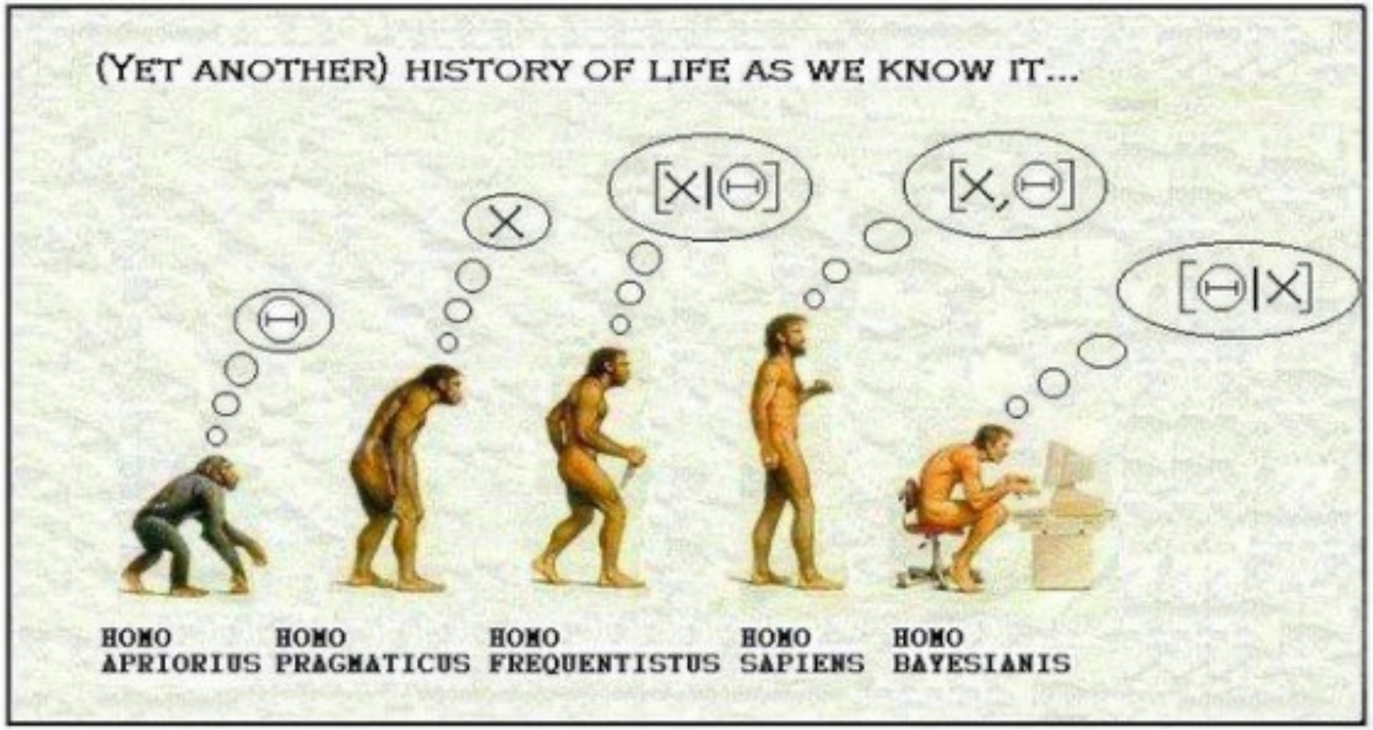
Regla de Bayes:

$$\mathrm{P}(X|Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{P(Y|X) * P(X)}{P(Y)}$$

NAIVE BAYES: TALLER

Descarguen los archivos 05-Taller NaiveBayes que se encuentra en la carpeta de la sesión 5

Desarrollen las partes 1 y 2 del taller de repaso, correspondientes al cálculo de probabilidades básicas y probabilidad condicional



(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...
[Θ]
[X]
[X|Θ]
[X,Θ]
[Θ|X]
HOMO APRIORIUS
HOMO PRAGMATICUS
HOMO FREQUENTISTUS
HOMO SAPIENS
HOMO BAYESIANIS

NAÏVE BAYES

PROBABILIDAD BAYESIANA

La estimación de la probabilidad de un evento, o un resultado potencial, debe basarse en la evidencia dada por múltiples ensayos u oportunidades para que ocurra el evento

Los métodos bayesianos proporcionan información sobre cómo la probabilidad de estos eventos puede ser estimada a partir de los datos observados

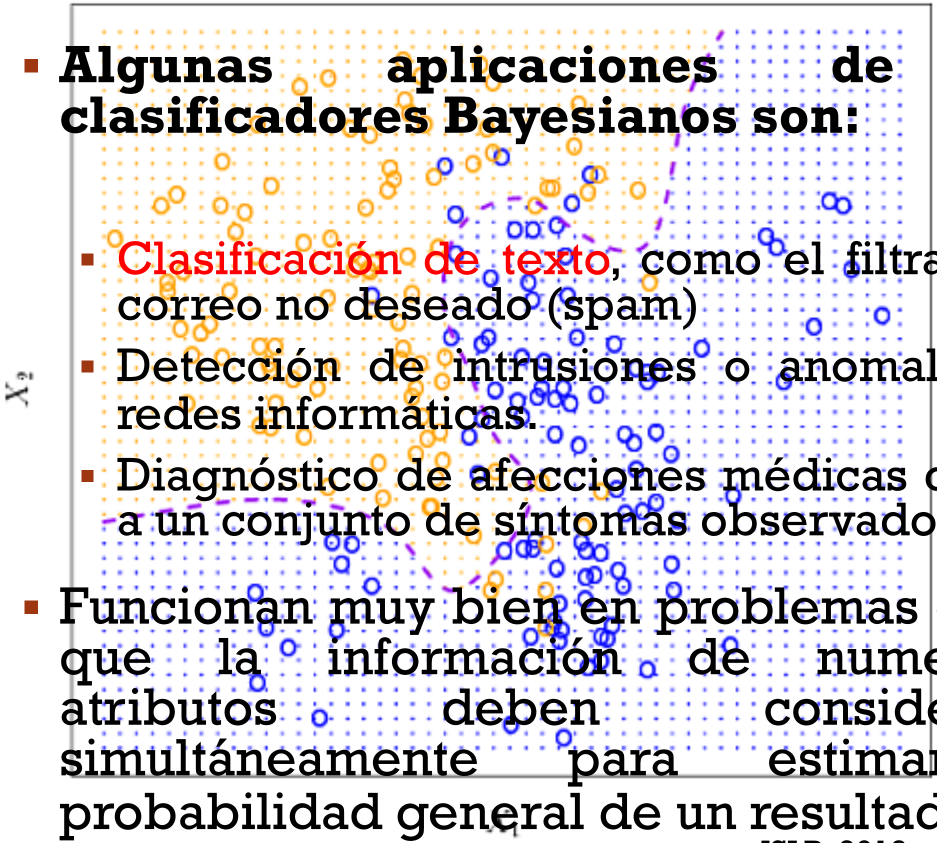
Los principios básicos de probabilidad se usan transversalmente en el algoritmo **Naïve Bayes**

CLASIFICADORES BAYESIANOS

- **Los clasificadores bayesianos** asignan cada observación a la clase j más probable, dados los valores observados de sus variables predictivas:

$$argmax_j \ p(Y = y_j | X = x_{observados})$$

- Si se conocen las distribuciones de probabilidad, el clasificador resultante da la frontera de separación óptima en términos de error
- No siempre se tienen las probabilidades condicionales necesarias.
- **Naïve Bayes** es un algoritmo basado en el Teorema de Bayes



- **Algunas aplicaciones de los clasificadores Bayesianos son:**
  - Clasificación de texto, como el filtrado de correo no deseado (spam)
  - Detección de intrusiones o anomalías en redes informáticas.
  - Diagnóstico de afecciones médicas debido a un conjunto de síntomas observados.
- Funcionan muy bien en problemas en los que la información de numerosos atributos deben considerarse simultáneamente para estimar la probabilidad general de un resultado

ISLR, 2013

UNIVERSIDAD ICESI

NAIVE BAYES (BAYES INGENUO)

Teorema de Bayes: $$\underbrace{p(y_j|x_1, x_2, \dots, x_n)}_{\text{Probabilidad a posteriori}} = \frac{p(y_j, x_1, x_2, \dots, x_n)}{p(x_1, x_2, \dots, x_n)} = \frac{\overbrace{p(y_j)}^{\text{Probabilidad a priori}} * \overbrace{p(x_1, x_2, \dots, x_n|y_j)}^{\text{Verosimilitud}}}{\underbrace{p(x_1, x_2, \dots, x_n)}_{\text{Verosimilitud marginal}}}$$

- El denominador es solo usado para propósitos de normalización (suma de probabilidades = 1)

$$p(x_1, x_2, \dots, x_n) = \sum_j p(y_j) * p(x_1, x_2, \dots, x_n|y_j)$$

- Por ello solo nos fijamos en el numerador:

$$p(y_j, x_1, x_2, \dots, x_n) = p(y_j) * p(x_1|y_j) * p(x_2|x_1, y_j) * p(x_3|x_2, x_1, y_j) * \dots * p(x_n|x_{1:n-1}, y_j)$$

- Si asumimos ingenuamente (**naïvely**) que todas las variables predictivas x_i son independientes condicionalmente con respecto a la clase y_j[1] entonces el numerador se simplifica a:

$$p(y_j) * p(x_1|y_j) * p(x_2|y_j) * p(x_3|y_j) * \dots * p(x_n|y_j)$$
$$= p(y_j) \prod_{i=1}^{n} p(x_i|y_j)$$

1 optimalidad de NB On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning. 1997; 29:103-130.

NAÏVE BAYES (BAYES INGENUO)

- La regla de clasificación es:

$$\operatorname{argmax}_j\ p(y_j)\prod_{i=1}^{n} p\left(x_i \middle| y_j\right)$$

- **Sólo necesitamos especificar :**
  - → Las probabilidades a priori de cada clase
  - → Las distribuciones de probabilidad de las variables predictivas para cada clase (condicionadas a la clase)

- Esta información se constituye en los **parámetros** del modelo, y en el caso de variables categóricas se obtienen a partir de tablas de frecuencias (conteos)

NAIVE BAYES

Ejemplo: Un banco quiere predecir si un cliente va a adquirir un CDT.

Creamos un clasificador Naïve Baye a partir de los datos históricos para calcular las probabilidades posteriores de cada clase: subscribed=yes and subscribed=no.

{Single, Married, Divorced}

{Subscribed Yes, Subscribed No}

$$p(y_j|x_1, \dots, x_n) = argmax_j \; p(\boxed{y_j}) \prod_{i=1}^{n} p(\boxed{x}|\boxed{y_j})$$

Marital	Subscribed=yes
Single	35%
Married	53%
Divorced	12%

Marital	Subscribed=no
Single	28%
Married	61%
Divorced	11%

Subscribed=yes	12%

Subscribed=no	88%

¿Debería el banco ofrecerle un CDT al cliente con la información siguiente?

Job=Management
Marital=Married
Education=Secondary
Default=no
Housing=yes
Loan=no
Contact=Cellular
Outcome=Success

Suponga que se disponen de las probabilidades condicionales para todas las variables predictivas (ya ilustradas para el estado civil "Marital")

UNIVERSIDAD ICESI

NAIVE BAYES

Ejemplo: Un banco quiere predecir si un cliente va a adquirir un CDT.

Creamos un clasificador Naïve Bayes a partir de los datos históricos para calcular las probabilidades posteriores para cada clase: subscribed=yes and subscribed=no.

$$p(y_j|x_1, \ldots, x_n) = argmax_j \; p(y_j) \prod_{i=1}^{n} p(x_i|y_j)$$

Marital	Subscribed=yes
Single	35%
Married	53%
Divorced	12%

Marital	Subscribed=no
Single	28%
Married	61%
Divorced	11%

Subscribed=yes	12%

Subscribed=no	88%

¿Debería el banco ofrecerle un CDT al cliente con la información siguiente?

	Subscribed=yes	Subscribed=no
Job=Management	22%	21%
Marital=Married	53%	61%
Education=Secondary	46%	51%
Default=no	99%	98%
Housing=yes	35%	57%
Loan=no	90%	85%
Contact=Cellular	85%	62%
Outcome=Success	15%	1%
Priors	12%	88%

Numerador	0,000255914	0,000169244
Proba posterior	60%	40%

NAÏVE BAYES (BAYES INGENUO)

- ¿Qué pasa si algunos de los valores de las variables predictivas tienen frecuencia nula con respecto a las categorías de la clase? ¿cuáles serían sus probabilidades aposteriori asociadas?
- Para evitar este problema, se utilizan métodos de **suavización.**
  - Por ejemplo, al contar las frecuencias de ocurrencia de cada valor se les agrega un valor pequeño, $\boldsymbol{\varepsilon > 0}$, evitando que alguna probabilidad sea cero:

$$P(casado|cliente\ potencial) = \frac{Conteo(casado,\ cliente\ potencial) + \varepsilon}{Conteo(cliente\ potencial) + N(x) * \varepsilon}$$

- El método de suavización de **Laplace** se aplica usualmente con $\varepsilon=1$, otro valor puede ser 1/n donde n es el número de datos de entrenamiento.

NAÏVE BAYES (BAYES INGENUO)

- Cuando las variables predictivas no son categóricas (e.g. numéricas), es necesario establecer una distribución de probabilidad:
  1. Se puede discretizar (en compartimentos) la variable convirtiéndola en categórica.
  2. Se establece una distribución de probabilidad empírica utilizando KNN,

$$P(Y = j|X = x_0) = \frac{1}{k}\sum_{i\in\aleph_0} I(y_i = j)$$

  3. Se supone que se trata de un tipo de distribución de probabilidad y se utiliza su función de densidad.
     - Por ejemplo, si se supone la variable sigue una **distribución normal** condicionada a la categoría objetivo, se puede calcular la media μ y desviación estándar σ a partir de los datos históricos, y utilizar la función de densidad:

$$P(edad|cliente\ potencial) = \frac{1}{\sigma_{edad|cliente}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{edad-\mu_{edad|cliente}}{\sigma_{edad|cliente}}\right)^2}$$

NAÏVE BAYES (BAYES INGENUO)

Pros:

- **Simple, rápido** y muy **efectivo**, permite atributos tanto categóricos como numéricos
- Estima efectivamente **las probabilidades condicionales** con respecto a los valores de la categoría objetivo
- Trabaja bien con atributos categóricos, con **valores faltantes** y con ruido
- Resistente al **overfitting**, sobretodo si se incluye un suavizador (e.g. Laplace)
- Trabaja bien con muestras de entrenamiento pequeñas y también con grandes

Contras:

- Sólo se puede utilizar para **clasificación**
- Se basa en **suposiciones** muy fuertes
- **Muy sensible** a atributos correlacionados (considera varias veces los mismos efectos)
- Las probabilidades estimadas son menos confiables que las clases predichas

NAIVE BAYES: TALLER

1. Continuando con el taller de Naive Bayes:

   Desarrollen la parte 3, aplicación de Bayes ingenuo para dos variables predictivas categóricas, y la parte 4, de aplicación de Bayes ingenuo para variables predictivas numéricas

EJEMPLO DE NAIVE BAYES

- 05-Naive-Bayes-Ejemplo
  - Desarrollo del Naïve Bayes desde cero
  - Naive_bayes de sklearn

TALLER: NAIVE BAYES, APLICACIÓN

- Ejecutar el cuaderno de Naive Bayes, que se encuentra especificado en el documento 05-01-NaiveBayes-Iris

PREGUNTAS CLAVE PARCIAL

1. ¿Qué es la Métodología ASUM-DM y en qué consiste?
2. ¿Cuáles son las características diferenciadoras entre modelos paramétricos y no paramétricos?
3. Explique en que consiste el aprendizaje supervisado y sus tareas.
4. Explique los conceptos de overfitting y underfitting (incluya los conceptos de sesgo y varianza).
5. Indique las características esenciales del K-NN y Naïve Bayes, además los casos en los que se pueden utilizar.
6. Explique que es el baseline, y cada una de las métricas de clasificación y regresión.
7. ¿Cuáles son las características diferenciadoras entre los distintos protocolos de evaluación?
8. Explique la matriz de confusión y su valor al momento de analizar los resultados de un modelo de clasificación.

REFERENCIAS

- *Introduction to Statistical Learning with Applications in R (ISLR), G. James, D. Witten, T. Hastie & R. Tibshirani, 2014*
- *Machine Learning with R*, Brett Lantz, Packt Publishing, 2015
- *Machine Learning,* Tom M. Mitchell, McGraw-Hill, 1997