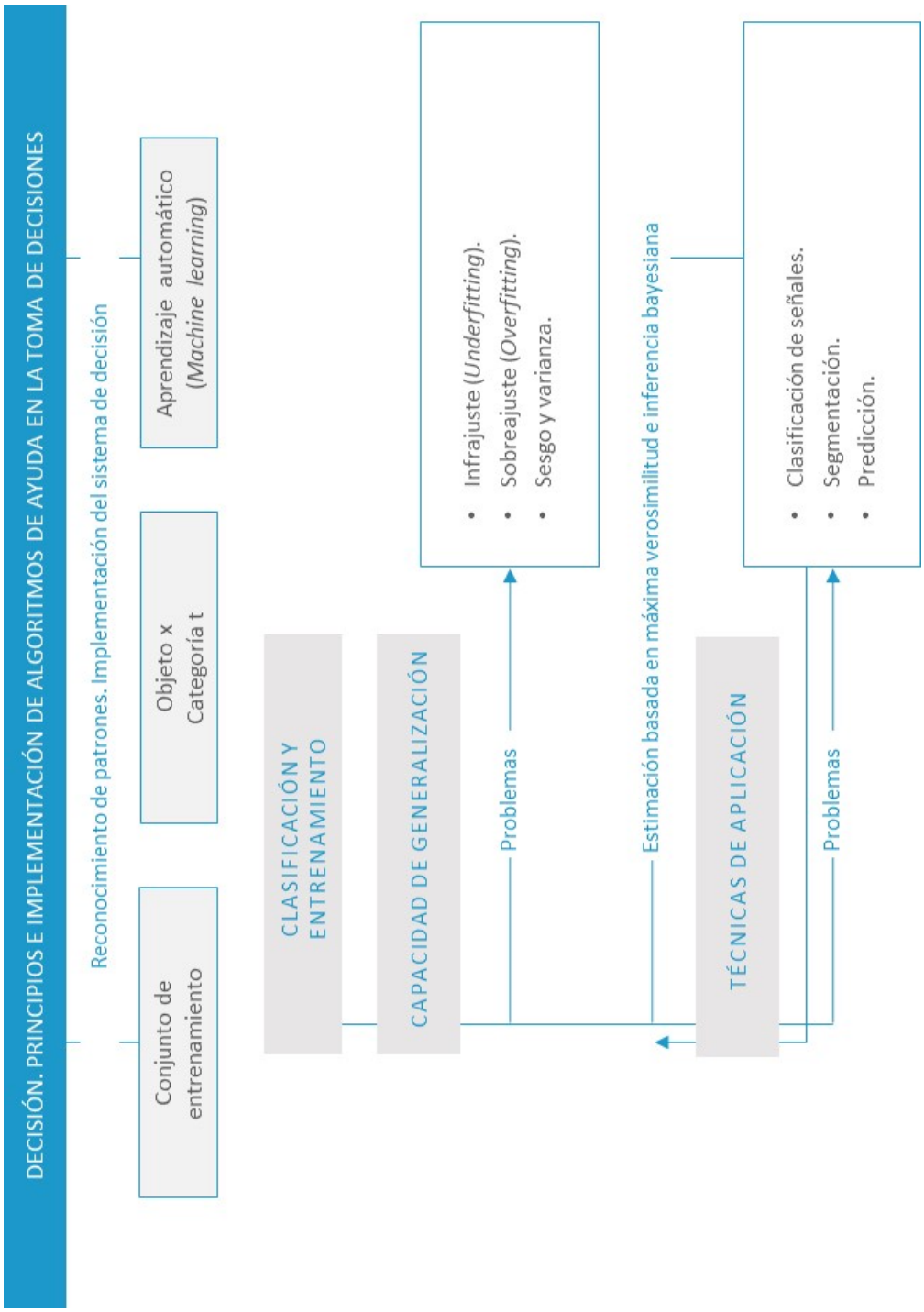
Visión Artificial

Tema 14. Decisión. Principios e implementación de algoritmos de ayuda en la toma de decisiones

Índice

Esquema



14.1. ¿Cómo estudiar este tema?

Para estudiar este tema deberás leer con atención las ideas clave que se desarrollan a continuación.

En la mayoría de las aplicaciones, el procesado completo de una señal requiere la interpretación automática de esta como etapa final. Por ejemplo, en el caso de una señal biomédica, el objetivo final irá dirigido a determinar si dicha señal denota la existencia o no de una patología. Podemos, asimismo, pensar en una aplicación basada en el manejo de señales de voz que nos permita identificar si la palabra o frase pronunciada forma parte de nuestro diccionario particular, que habrá sido definido previamente. En imágenes es habitual tener que identificar determinados objetos, por ejemplo, conocer si corresponde o no a una persona; por lo que tras reducir el ruido de la imagen, incrementar el contraste de esta o llevar a cabo una segmentación, será necesario clasificar los elementos resultantes.

Todas estas aplicaciones precisan como elemento culminante una etapa de **reconocimiento de patrones.** Esta será específica de la aplicación en cuestión y, por tanto, su construcción se adaptará a la tarea concreta que deberá llevar a cabo. Así, el sistema de reconocimiento de patrones para la identificación de palabras es diferente al empleado para hallar automáticamente una persona en una imagen. Sin embargo, la metodología para su construcción es común en ambos casos.

Dicha **metodología** se basa en la disponibilidad de un conjunto amplio de ejemplos del problema a resolver. Por ejemplo, a fin de implementar un sistema de reconocimiento de voz, este conjunto deberá incluir tanto palabras que no están en nuestro diccionario como las que sí lo están. O en el caso de las imágenes, instantáneas de figuras que no se corresponden con una persona frente a las que sí lo son. A partir de estos ejemplos, es posible inferir qué propiedades diferencian unos casos de otros, resultando en un **modelo.**

En este punto, cabe destacar que cada una de estas tareas es realizada por nuestro cerebro, que es capaz de identificar una palabra conocida o si el elemento de una imagen es una persona. En él se llevan a cabo las etapas previas de tratamiento de señal que se han visto en la asignatura: captura, eliminación de ruido, segmentación y extracción de características a partir de la fuente de información original.

La etapa inmediatamente anterior al sistema de reconocimiento de patrones es la **extracción de características** y sobre la que hemos tratado en temas previos. En esta etapa, una vez que tenemos identificado el objeto a interpretar (una señal o segmento de la misma, una imagen, una textura, una silueta, etc.), se busca sintetizar la información de dicho objeto en un conjunto de variables. Este ejercicio de síntesis conlleva la reducción de la dimensionalidad de los datos de entrada, de una imagen o señal formados por un número determinado de píxeles o puntos, pasamos a un vector de características de longitud notablemente menor.

La motivación principal de la etapa de extracción de características es optimizar, mediante esta síntesis, la implementación del sistema de reconocimiento de patrones posterior. Para ello, además de permitir trabajar en un espacio de dimensión menor, es de vital importancia que las características extraídas definan propiedades significativas de los objetos en el contexto del problema a resolver. Por ejemplo, el valor máximo de una señal de voz puede no tener relevancia si lo que se pretende identificar es si ha sido emitida por un hombre o una mujer, pero sí si se pretende

inferir si el emisor de la señal muestra un estado de tranquilidad o no. Es decir, la etapa de extracción de características ha de minimizar la pérdida de información resultante de la reducción de dimensionalidad llevada a cabo, reteniendo aquellas propiedades más relevantes de la fuente de información.

La siguiente imagen recoge, en forma de esquema, los principales conceptos de este tema.

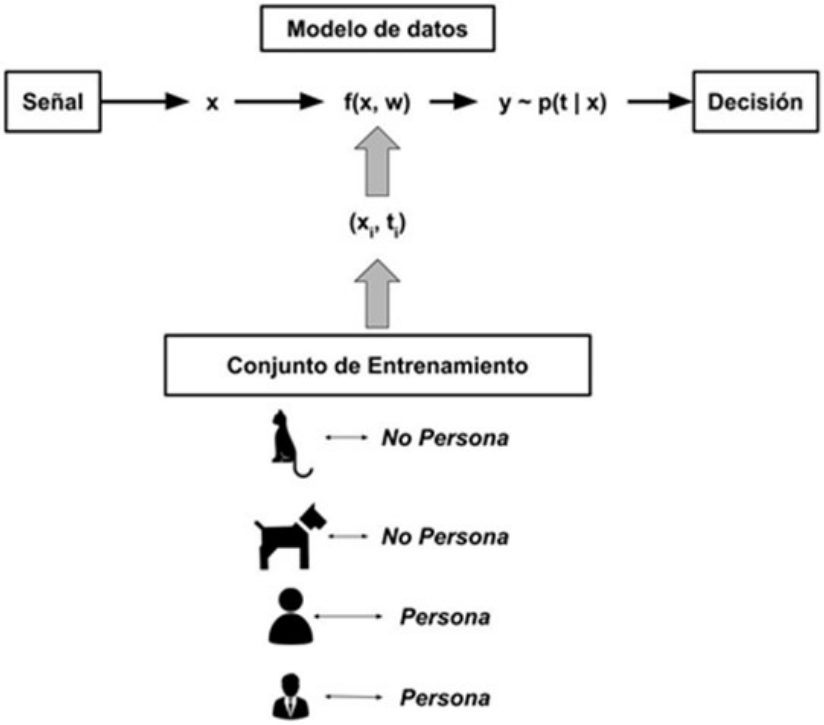


Figura 1. Metodología para la implementación de un sistema de decisión.

14.2. Clasificación y reconocimiento de patrones

El sistema de reconocimiento de patrones es el **modelo** resultante de inferir las diferencias existentes entre varias categorías de objetos. Dicho modelo, como se expresaba al comienzo de este tema, se obtiene a partir de un conjunto de ejemplos de todas las categorías involucradas en nuestra aplicación. Finalmente, nuestro modelo se deriva de un proceso de optimización guiado por un algoritmo y una métrica de evaluación a maximizar/minimizar. Este algoritmo suele ser de carácter iterativo, de manera que se va dando forma al modelo mediante la observación repetida de los diferentes ejemplos.

Como se ha comentado, la implementación del sistema de decisión se realiza a partir de:

- Un conjunto de ejemplos D, referido comúnmente como **conjunto de entrenamiento.**
- Cada uno de los ejemplos en este conjunto viene dado por el par formado por:
  - Un objeto x.
  - Y la categoría a la que pertenece t.
- Este escenario se conoce en el ámbito del aprendizaje automático *(machine learning)* como aprendizaje supervisado, ya que se sabe *a priori* el valor de la variable objetivo para los datos de entrenamiento disponibles.

El conjunto D puede representarse, por tanto, de la siguiente forma:

$$D = \{(x_i, t_i)\}_{i=1,\ldots,M}$$

Donde M es el número total de ejemplos en el conjunto de entrenamiento.

La pregunta es cómo tomar decisiones, minimizando el error en la decisión, ante objetos x que no están en D y para los que desconocemos la categoría a la que pertenecen.

Estadísticamente, la herramienta que proporciona una descripción completa de las muestras en D es la función de densidad de probabilidad conjunta $p(x,t)$, que puede expresarse como:

$$p(x,t) = p(t|x)\, p(x)$$

Donde:

- $p(t|x)$ es la función de densidad de probabilidad de la variable t dado x.
- $p(x)$ es la función de densidad de probabilidad de x.

En este punto, puede observarse que $p(t|x)$ es la función a modelar a fin de hacer predicciones sobre t a partir de un valor observado de x. Es decir, esta será la **función a modelar por nuestro sistema de decisión.**

Clasificación

Cuando la variable t es categórica, el problema se corresponde con una tarea de clasificación.

Matemáticamente consiste en determinar, para una muestra x, a cuál de las C categorías posibles $C_1, C_2, \ldots, C_k$ pertenece. El objeto de entrada queda

descrito por un conjunto de características o descriptores cuantitativos, por lo que x será un vector de la forma $x = [x_1, x_2, \ldots, x_n]$, que representa un punto en el espacio de entrada n-dimensional.

El **clasificador** viene dado por la función *f* que mapea el espacio de entrada R^n en la variable categórica t:

$$f: R^n \rightarrow t \in \{C_1, C_2, \ldots, C_k\}$$

La salida del clasificador puede expresarse como $y = f(x, w)$, donde se hace explícita la dependencia de la función de mapeo f con el vector de entrada x y un conjunto de parámetros adaptativos w.

La aproximación estadística al problema de clasificación, es decir, para la estimación de la función f, asume que las muestras de x que corresponden a la categoría C_i han sido generadas de acuerdo a la función de densidad de probabilidad $p(x|C_i)$.

En este escenario, la **regla de decisión de Bayes** determina cómo ha de llevarse a cabo la decisión a fin de minimizar el riesgo en la predicción. De acuerdo a esta regla, un objeto x debe asignarse a la categoría C_i para la que el riesgo condicionado $R(C_i|x)$ es mínimo.

Este riesgo se define como sigue:

$$R(C_i|x) = \sum_{j=1}^{k} L(C_i, C_j)\, p(C_j|x)$$

Donde:

- Donde la función $L(C_i, C_j)$ cuantifica el **coste del error** cuando se decide C_i siendo C_j la verdadera categoría a la que pertenece la muestra x.

Ideas clave

- $p(C_j|x)$ es la probabilidad *a posteriori* de la categoría C_j . Se conoce también como función de pérdida. Normalmente, se emplea una función de pérdida 0/1 para la implementación del clasificador, que toma un valor:
  - 0 si $i = j$.
  - 1 si $i \neq j$.

El resultado de utilizar esta función de pérdidas es la denominada como **regla del máximo *a posteriori*** (MAP, del inglés *maximum a posteriori).* Supone la forma más conocida de la regla de decisión de Bayes. Una muestra x se asigna a la categoría C_i si se cumple la siguiente condición:

$$p(C_i|x) > p\left(C_j|x\right), i \neq j$$

Es decir, la decisión será tomar aquella categoría más probable una vez que el valor de x es conocido.

La estrategia descrita al problema de reconocimiento de patrones y, concretamente, a la clasificación representa una aproximación estadística al mismo. En ella, se pretende obtener una estimación de la función $p(C_i|t)$ a fin de poder tomar decisiones de acuerdo a la regla de Bayes. Muchos de los **algoritmos para la implementación de modelos de datos** se basan en esta aproximación, es el caso de la regresión logística o las redes neuronales MLP y RBF. Asimismo, en otros modelos estadísticos se estima la función de densidad condicionada $p(x|C_i)$ para obtener $p(C_i|x)$ mediante el teorema de Bayes:

$$p(C_i|x) = p(x|C_i)\, p(C_i) \,/ p(x)$$

Algunos algoritmos que siguen esta estrategia son K-vecinos próximos, las redes neuronales probabilísticas, los clasificadores gaussianos (discriminante lineal o cuadrático) o el Naïve Bayes.

Fuera de este marco puramente estadístico, podemos encontrar métodos de clasificación como los árboles de decisión o las máquinas de vectores soporte. En ambos casos, no se persigue una estimación directa de las funciones de densidad de probabilidad mencionadas previamente.

Entrenamiento

En principio, la forma de la función $f(x,w)$ es desconocida y se requiere de un conjunto de ejemplos D para su estimación. Por tanto, la función f vendrá dada por una expresión matemática que contiene un número determinado de parámetros ajustables w. Esta expresión está determinada por el conjunto de ejemplos entrada-salida en D mediante el proceso de aprendizaje o entrenamiento. Por este motivo, el conjunto de ejemplos de muestras de entrada y su correspondiente categoría se denomina conjunto de entrenamiento.

El proceso de entrenamiento conlleva la minimización de una función de error que cuantifica la distancia o grado de disparidad entre el valor de salida de nuestro clasificador (y) y su valor objetivo correspondiente (t) para un objeto dado (x). En este caso, se habla de entrenamiento supervisado, ya que cada muestra x está asociada a un valor objetivo t.

Ideas clave

La finalidad del proceso de entrenamiento es construir un modelo estadístico del proceso que da lugar a las muestras x. Es decir, no se busca una función $f(x,w)$ cuyo valor coincida exactamente con la variable t para cada par (x,t), sino que dicha función se aproxime a $p(t|x)$. Una estimación plausible de esta función de densidad de probabilidad resultará en un modelo con una gran **capacidad de generalización;** este concepto se refiere a la capacidad del modelo de producir respuestas precisas ante muestras de entrada que no se han procesado previamente. Fundamentalmente, esta capacidad de generalización depende de tres factores:

El tamaño del conjunto de entrenamiento y su representatividad del problema. Será necesario contar con un número elevado de ejemplos diferentes entre sí para lograr una buena descripción del problema. Cada ejemplo en D es, al fin y al cabo, una muestra de la función $p(x,t)$, por lo que su estimación será más precisa si se dispone de un mayor número de muestras.

La capacidad de adaptación (flexibilidad) del modelo. Cuanto mayor sea el número de funciones $f(x,w)$ que pueden definirse mediante el ajuste de w, mayor es la posibilidad de encontrar una solución que minimice la distancia respecto de la función objetivo $p(t|x)$, resultando en una mejor capacidad de generalización. Este número de funciones a las que el modelo puede dar lugar viene dado por la cantidad de parámetros w ajustables en el mismo. Así, cuantos más parámetros se deban optimizar durante el entrenamiento, el modelo tendrá una capacidad de adaptación mayor a las muestras de $p(t|x)$ recogidas en D.

La complejidad del problema de clasificación. Un problema de clasificación de mayor dificultad para modelar vendrá dado por una función $p(t|x)$ más compleja. La siguiente imagen (figura 2) muestra un ejemplo en el que se aprecian dos conjuntos de muestras (x,t):

- D1 (izquierda) y D2 (derecha) describen dos problemas de clasificación diferentes.
- Como se aprecia, el problema definido por D2 es de mayor dificultad, pues requiere una frontera de decisión en el espacio bidimensional de entrada de la forma:

$$(x_1 - k_1)^2 + (x_2 - k_2)^2 = k_3$$

Donde k_1 ,$k_2 y k_3$ son constantes y los objetos x de entrada quedan definidos por los valores de las variables x_1 y x_2 .

- Mientras tanto, el problema descrito por el conjunto de muestras D1 requiere una frontera de decisión de la siguiente forma:

$$x_1 = k_1 x_2 + k_2$$

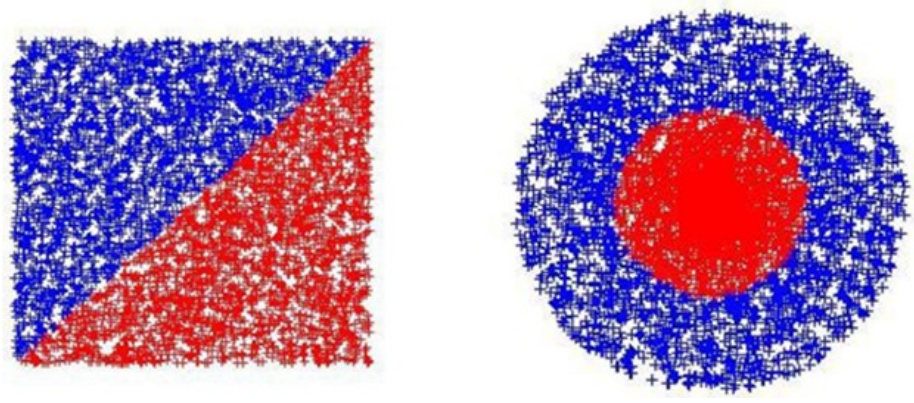


Figura 2. Ejemplo de dos problemas de clasificación diferentes.

Como comentábamos, los datos de la imagen de la derecha reflejan una mayor complejidad de la función $p(t|x)$ que los que generó respecto a los observados en la imagen de la izquierda.

Mientras la dificultad del problema de clasificación resulta un factor fuera de control para el usuario, los dos primeros factores están estrechamente relacionados. Un **modelo rígido o poco flexible,** con un número reducido de parámetros ajustables durante el entrenamiento, puede no ajustarse a las muestras de la función a aproximar. Lo contrario sucede con un modelo excesivamente **flexible y maleable,** pues será capaz de ajustar perfectamente los pares en D.

En ambas situaciones, la capacidad de generalización del modelo resultante será pobre.

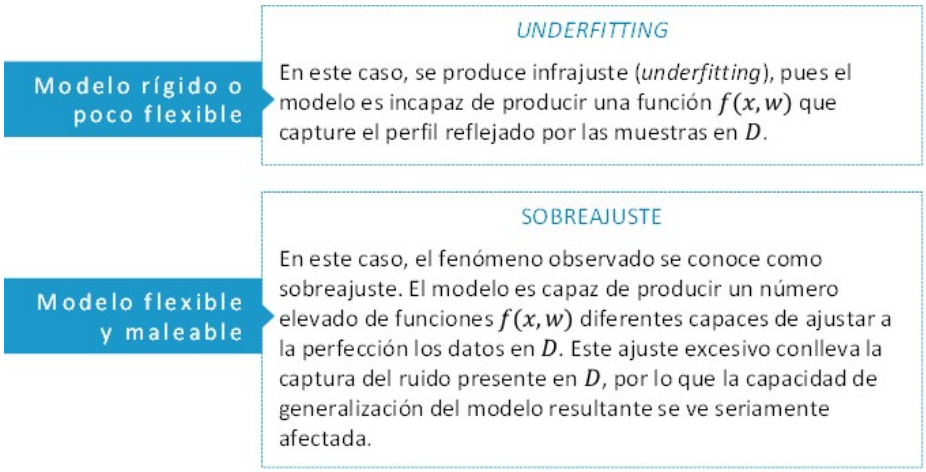


Figura 3. Problemas de infrajuste o sobreajuste en la capacidad de generalización del modelo.

Existen diferentes estrategias para paliar ambos efectos:

- En el caso del *underfitting,* la solución más efectiva es emplear modelos con un mayor grado de flexibilidad, de forma que sean capaces de generar funciones $f(x,w)$ de un perfil más complejo.
- En el caso del *overfitting* (sobreajuste), puede prevenirse mediante la utilización de un conjunto de entrenamiento más amplio. Sin embargo, esto no es posible en la mayoría de los escenarios reales. Por tanto, se suele optar por otras soluciones más fácilmente alcanzables como la reducción de la dimensión del espacio de entrada o la combinación de modelos diferentes entre sí, técnicas conocidas como *ensembles* de las que *bagging* y *gradient boosting* son las más comúnmente empleadas.

Sesgo y varianza

Los conceptos de infrajuste y sobreajuste mencionados se encuentran estrechamente relacionados con el sesgo y la varianza de un modelo de datos. Estos representan dos componentes distintas del error en el que incurre un modelo de datos. Antes de definirlas, la siguiente imagen ilustra la idea detrás de ambas componentes.

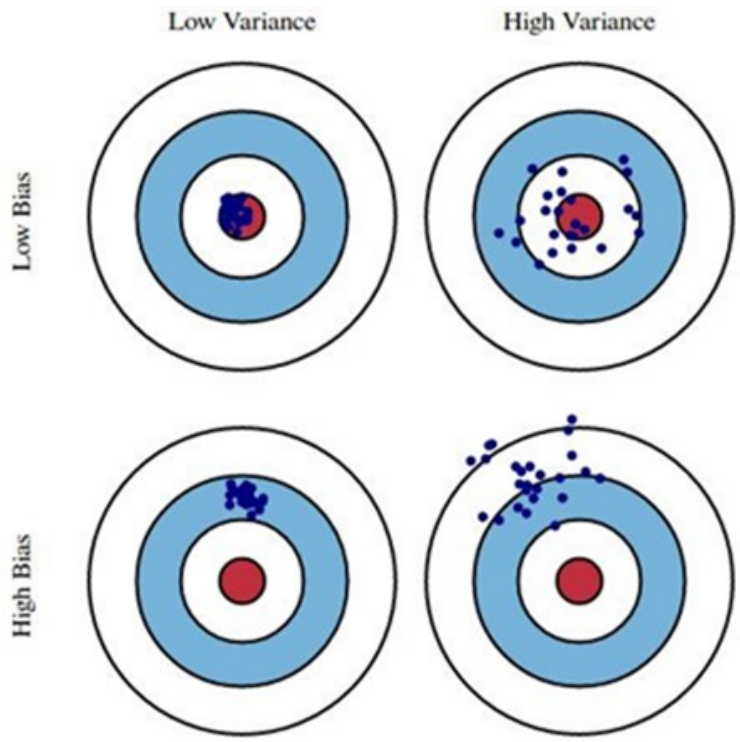


Figura 4. Ilustración de los conceptos de sesgo y varianza como fuentes de error en un modelo de datos.
Fuente: https://www.kdnuggets.com/2016/08/bias-variance-tradeoff-overview.html

- Si atendemos en primer lugar al sesgo, vemos que, cuando este es elevado en un modelo, el error cometido por este se debe a una tendencia clara en la predicción. En la imagen anterior, el sesgo del modelo se refleja en una tendencia a estimar por encima del objetivo. Esta tendencia desaparece cuando el modelo tiene un sesgo bajo.

- Por otro lado, la varianza se manifiesta en una distribución más amplia de las predicciones generadas por el modelo a pesar de que el objetivo, como en el ejemplo, no cambia. Así, cuando la varianza del modelo es elevada, hay mayor grado de incertidumbre sobre la salida de este.

Por tanto, tras este primer análisis, podemos definir sesgo y varianza:

El **sesgo** obedece a la componente sistemática del error. Dada una entrada x , se computa como la diferencia entre el valor medio de las predicciones de los modelos derivados de diferentes conjuntos de entrenamiento D y el valor objetivo t asociado a x .

$$Bias^2 = [E_D \{y(x)\} - t \vee x]^2$$

El error asociado a un sesgo elevado es propio de modelos simples, con poca flexibilidad o capacidad de ajuste. Estos modelos son incapaces de aproximar funciones complejas y tienden a producir infraestimación o sobreestimación. Esta situación refleja infrajuste de los datos en D .

La **varianza** es la componente del error derivada de la dependencia del modelo resultante $f(x,w)$ con el conjunto de entrenamiento D . Es decir, este término refleja que las funciones $f(x,w)$ derivadas del proceso de entrenamiento difieren entre sí cuando los conjuntos de entrenamiento empleados son distintos.

$$Varianza = E_D [\{y(x) - E_D \{y(x)\}\}]^2$$

El error derivado de la varianza es característico de modelos con una gran capacidad de adaptación a los datos. Potencialmente, estos modelos son capaces de definir diferentes funciones que ajustan los datos en D , capturando incluso el ruido presente en este conjunto en forma de anomalías. Se produce, por tanto, el sobreajuste de los datos de entrenamiento. Un mayor número de ejemplos en D contribuiría a limitar el espacio de funciones que el modelo sería capaz de definir para el ajuste de D y, por tanto, se reduciría su varianza.

Estimación basada en máxima verosimilitud e inferencia bayesiana

Tras el análisis de las fuentes de error en un modelo de datos, retomamos el proceso de entrenamiento. Como se ha indicado, en el mismo se tiene como objetivo modelar el generador de los datos en D, que queda completamente descrito por la función $p(x,t)$.

A fin de estimar el valor de t para un x dado, la función e interés será $p(t \vee x)$. Para incluir explícitamente la dependencia del modelo en la estimación, puede expresarse $p(t \vee x,w)$, donde w refleja el conjunto de parámetros del modelo ajustables durante el proceso de entrenamiento. Estadísticamente, esto puede obedecer a dos principios: regla de máxima verosimilitud o inferencia bayesiana.

El principio de máxima verosimilitud ajusta los pesos w de tal forma que, dado estos, se maximice la probabilidad de observar los datos en D.

Si asumimos independencia entre las muestras en D, la probabilidad de observar este conjunto de datos según el modelo definido por w puede expresarse de la siguiente forma:

$$p(D \vee w) = \prod_{i=1}^{M} p(x_i,t_i \vee w) = \prod_{i=1}^{M} p(t_i|x_i,w) \equiv L_D(w)$$

Donde el término $L(w)$ se identifica como la probabilidad asociada al conjunto D y depende del modelo w.

Los pesos de w se escogen de forma que se maximice esta función, es decir, se elige el modelo que hace más probable dar lugar al conjunto de datos. En la práctica, se lleva a cabo la operación equivalente que consiste en minimizar el logaritmo negativo de L_D . La **función de error** resultante se obtiene como:

$$E_D = -\log L_D = -\sum_{i=1}^{M} \log \left[p\left(t_i|x_i,w\right)\right] - \sum_{i=1}^{M} \log \left[p\left(x_i\right)\right]$$

Dado que el segundo término de E_D no depende de w , la función de error resultante viene dada por el primer sumando de la ecuación. En el caso de un problema de clasificación con dos categorías diferentes $\{C1,C2\}$, de forma que $t=1$ para C1 y $t=0$ para C2, la **función aproximada por la salida del modelo** puede escribirse de la siguiente forma:

$$p\left(t|x,w\right) = y^t\left(x,w\right)\left[1-y\left(x,w\right)\right]^{1-t}$$

Sustituyendo esta expresión en la ecuación del error de acuerdo al principio de máxima verosimilitud, obtenemos la siguiente ecuación que se conoce como **función de entropía cruzada:**

$$E_D = \sum_{i=1}^{M} \left\{t_i \log\left(y_i\right) + \left(1-t_i\right)\log\left(1-y_i\right)\right\}$$

Sin embargo, el modelo que asegura una mayor probabilidad de observar D mediante la minimización del error anterior, no asegura la mejor capacidad de generalización. Este hecho se ha visto previamente con el problema de sobreajuste, común en modelos con un alto grado de varianza. La aproximación bayesiana para el ajuste de w considera este conjunto de parámetros como una variable aleatoria, de forma que esté caracterizada por una función de densidad de probabilidad $p\left(w\right)$. Una vez que el conjunto D es observado, la probabilidad asociada a w cambia y puede ser obtenida a través del **teorema de Bayes:**

$$p(w|D) = \frac{p(D|w)\,p(w)}{p(x)}$$

La estimación de $p(w|D)$ permite realizar predicciones teniendo en cuenta todos los posibles valores de la variable w, es decir, todos los modelos potencialmente obtenibles a partir de D tras la observación de este conjunto. Así, el **valor de salida del modelo** vendría dado por la siguiente expresión:

$$p(t|x,D) = \int p(t|x,w)\,p(w|D)\,dw$$

Donde $p(t|x,w)$ sería el valor de salida de un modelo w concreto para la entrada x. Por tanto, el valor final de salida es un promediado de las salidas de todos los modelos w observables, ponderados por la probabilidad de cada uno de ellos tras la observación de D, que viene dada por el término $p(w|D)$.

14.3. Aplicación de técnicas machine learning al procesado de señales

El punto anterior recoge los principios elementales en la implementación de un sistema de decisión o reconocimiento de patrones. Estos principios se han presentado en un marco estadístico que permite la utilización de herramientas matemáticas para la obtención de una solución al problema. A continuación, se indica cómo las técnicas derivadas de este marco pueden aplicarse a señales para la implementación de señales de decisión.

Comenzaremos con los problemas más habituales en el ámbito del tratamiento de señales que precisan de la implementación de un sistema de decisión.

Clasificación de señales

Una de las aplicaciones más comunes de los sistemas de decisión se encuentra en la identificación automática de la categoría de una señal, tanto si pensamos en una función unidimensional de la variable tiempo como en una imagen. En este escenario, puede que nos encontremos en una de las siguientes situaciones:

- El problema de clasificación no es sencillo de resolver por el ser humano mediante la observación de las señales. En este caso, precisamos de herramientas avanzadas que capturen información de la señal que nosotros no somos capaces de extraer.
- El ser humano es capaz de llevar a cabo la clasificación que se persigue, pero se trata de una tarea que precisa de elevados recursos (materiales, tiempo, profesionales involucrados...). Por tanto, en esta situación el objetivo es **automatizar el procedimiento.**

El problema de clasificación de señales debe acotarse a un escenario concreto, previamente definido, a fin de obtener un sistema de decisión efectivo. Por ejemplo, en el caso de una imagen, el propósito de la tarea de clasificación se ceñiría a identificar si la figura que aparece en ella se corresponde o no con una persona. La interpretación automática de qué ser vivo o cosa es dicha figura, de entre todas las posibilidades existentes, es un problema excesivamente abierto y complejo de abordar. Otro caso típico en sistemas de ayuda a la decisión basado en imágenes se tiene en un contexto médico. Así, podemos pensar en la implementación de un clasificador que permita identificar automáticamente tejido pulmonar dañado frente a tejido sano sin lesiones relevantes.

De forma similar, en el caso de las **señales en el dominio del tiempo,** es habitual enfrentar problemas en los que el propósito es separar registros patológicos de aquellos correspondientes a un paciente sano. Suele ser común el procesado de señales como el electrocardiograma (ECG), el electroencefalograma (EEG) o la saturación de oxígeno (SaO).

Por último, las **herramientas de reconocimiento de voz** implementan también clasificadores de señales. Una herramienta relativamente sencilla sería un sistema capaz de distinguir si la voz percibida se corresponde con un hombre o una mujer. Las máquinas para la interpretación automática del habla poseen mayor complejidad. El sistema dispone de un léxico o conjunto de palabras con las que ha sido entrenado y, a partir de la señal de audio capturada, se busca cada una de las palabras en este léxico para poder interpretar el mensaje completo.

Segmentación

En el contexto del procesado de imágenes, los sistemas de decisión pueden emplearse para la identificación de diferentes regiones y el cálculo de las fronteras entre estas. Para ellos, es común emplear filtros espaciales que, centrados en un píxel y con un tamaño determinado, clasifican el píxel central asignándole una de las

posibles categorías. Si estas han sido previamente definidas, estamos ante un **problema supervisado,** por lo que el clasificador será entrenado a partir de un conjunto de ejemplos previamente etiquetados, tal y como hemos supuesto desde el inicio en este tema.

El **problema no supervisado** en el que no disponemos de un conjunto de ejemplos y hemos de inferir las diferentes categorías existentes, requiere de la utilización de técnicas de *clustering.* La figura 5 muestra un ejemplo de un problema de clasificación de regiones en una imagen.

Figura 5. Segmentación de regiones en una imagen. Fuente: http://web.eecs.umich.edu/~silvio/teaching/lectures/Vision%20Research%20Lab%20-%20A%20Multiresolution%20Approach%20to%20Image%20Segmentation%20Based%20on%20EdgeFlow.htm

Predicción

El trabajo con señales unidimensionales que son función del tiempo presenta la necesidad de estimar valores futuros de la señal a partir de muestras históricas. Como un ejemplo claro puede pensarse en cotizaciones bursátiles. El sistema de decisión consistiría en un clasificador capaz de capturar si la cotización del día siguiente será mayor o menor respecto a la del día anterior, basándose en muestras del histórico de la serie.

En los diferentes escenarios descritos, la **metodología para la implementación** del clasificador es común, tal y como se describía en el comienzo de este tema:

- Identificación del objeto a clasificar: píxel de una imagen, región de una imagen, segmento de señal, etc.
- Preprocesado de la fuente de información: eliminación de *outliers* y ruido, detección de bordes, selección de altas frecuencias, etc.
- Caracterización del objeto: definición del conjunto de atributos que describen el objeto sobre el que se va a decidir; este proceso se trata de la construcción del vector x de características a partir de las técnicas descritas en temas anteriores.
- Definición del problema a modelar: definición de la variable objetivo t : clasificación de una señal de ECG como patológica o no; identificación de una palabra en un diccionario; figura de una persona o no, etc.
- Construcción de un conjunto de ejemplos: conjunto de entrenamiento D formado por pares (x_i,t_i) .
- Entrenamiento y test: implementación del sistema de decisión y evaluación de su rendimiento sobre un conjunto de muestras no empleadas para el ajuste del modelo.

Aprende Machine Learning con Andrew Ng

Andrew Ng. (s.f.). *Courses.* http://www.andrewng.org/courses/

Página web de Andrew Ng: jefe científico de Baidu, cofundador de Coursera y profesor adjunto en la Universidad de Stanford. En su web puedes encontrar un listado de cursos sobre *Machine Learning* que él mismo imparte.

Kaggle

Página de Kaggle https://www.kaggle.com/

Página web dedicada al aprendizaje y, divulgación de todo lo relacionado con la Ciencia de datos *(Data Science)* y *Machine Learning.* Incluso tienen buscador de empleo.

Bibliografía

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Clarendon Press.

Fukunaga, K. (2013). *Introduction to statistical pattern recognition*. Cambridge: Academic Press.

Jain, A. K., Duin, R. P. W. y Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence, 22*(1), 4-37. doi: 10.1109/34.824819

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill Education.

Webb, A. R. (2003). Statistical pattern recognition. Hoboken: John Wiley & Sons. Recuperado de http://onlinelibrary.wiley.com/book/10.1002/0470854774

1. ¿En qué punto del proceso de tratamiento de una fuente de información nos encontramos la etapa de decisión?

A. Primer paso del proceso.

B. Última etapa del proceso.

C. Antes de eliminar el ruido de la señal.

D. Entre la identificación de *outliers* y la extracción de características.

2. ¿Por qué incluimos un elemento de decisión automática en nuestro sistema de procesado?

A. Necesitamos asignar nuestra señal o una parte de ella a una categoría o grupo de entre varios posibles, automatizando la interpretación de la información

B. Supone una parte más para la eliminación del ruido

C. Ninguna respuesta es correcta

D. Permite realzar los elementos característicos de la señal para su posterior procesado.

3. ¿Qué regla permite llevar a cabo la decisión minimizando el coste del error?

A. Teorema de Tales.

B. Regla de decisión estadística

C. Principio de máxima verosimilitud.

D. Regla de decisión de Bayes.

4. En el contexto de la implementación de un sistema automático de decisión, ¿qué es un conjunto de entrenamientos?

A. Datos representativos del problema a resolver.

B. Fuente de información a partir de la cual se implementa un modelo.

C. Conjunto de ejemplos correspondientes a muestras de entrada y su salida o decisión esperada.

D. Todas las respuestas son correctas.

5. De acuerdo al marco estadístico para la implementación de sistemas de decisión, si nuestro conjunto de entrenamiento se denota como $D = \{(xi, ti)\}$, ¿qué función caracteriza por completo al generador de información del que se deriva?

A. La función de densidad de probabilidad de x.

B. La función de densidad de probabilidad de t.

C. La función de densidad de probabilidad conjunta de x y t.

D. Todas son correctas.

6. A fin de tomar decisiones precisas a partir del valor de x, ¿qué función es la que debe modelar nuestro sistema de decisión?

A. La función de densidad de probabilidad de x.

B. La función de densidad de probabilidad de t.

C. La función de densidad de probabilidad conjunta de xt.

D. La función de densidad de probabilidad de condicionada a x.

7. El teorema de Bayes:

A. Permite obtener la probabilidad de pertenencia a una categoría dada la muestra de entrada x a partir de la probabilidad de observar x en cada una categoría la probabilidad *a priori* asociada a cada categoría y la probabilidad de observar x.

B. Permite obtener la función de densidad de probabilidad de x condicionada a t a partir de las probabilidades *a priori* de cada categoría

C. Permite relacionar categorías con muestras x a partir únicamente de la probabilidad de observar la muestra x.

D. Permite tomar decisiones minimizando el riesgo de error.

8. En el proceso de entrenamiento:

A. Se estima la función de densidad de probabilidad de

x .

B. Se ajustan los parámetros del modelo a fin de aproximar la función de densidad

$p(x,t)$ que dio lugar a las muestras en D.

C. Se maximiza la función de error que cuantifica la disparidad entre

x y

t .

D. Se optimiza la regla de decisión de Bayes.

9. ¿Cuáles son las componentes del error de un sistema de decisión derivado de un proceso de entrenamiento?

A. El sesgo, que representa el error sistemático de nuestro clasificador.

B. La varianza, que representa la dependencia del modelo resultante con el conjunto de entrenamiento empleado.

C. Sesgo y varianza, que son componentes complementarias.

D. Sesgo y varianza, que pueden ser minimizadas si se reduce el número de muestras en el conjunto de entrenamiento.

10. ¿Cuándo es más probable que se produzca *overfitting?*

A. Si nuestro modelo tiene un número muy reducido de parámetros ajustables w.

B. Si nuestro modelo tiene un número muy amplio de parámetros ajustables y podemos tener tantas muestras como queramos en el conjunto de entrenamiento.

C. Si la varianza del modelo es elevada.

D. Si el espacio de entrada es bidimensional y disponemos de miles de muestras diferentes en el conjunto de entrenamiento.