

Introducción al Aprendizaje automático

Dra Ana Georgina Flesia

Optativa Ciencias de la Computación

FaMAF-UNC

Oficina 370

georgina.flesia@unc.edu.ar

2020

Aprendizaje

Herbert Simon

Aprendizaje es un proceso por el cual un sistema mejora su desempeño usando experiencia previa.

Objetivo general del área

Estudiar la forma de construir programas de computación que mejoren **automáticamente** con nuevas experiencias

Aprendizaje

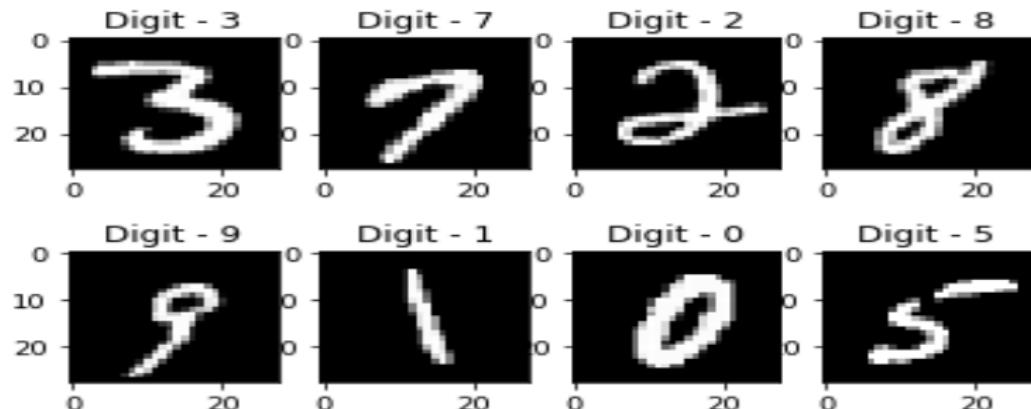
Tom Mitchell (1997)

Una máquina aprende una particular tarea **T** considerando las experiencias de tipo **E** respecto de una medida de desempeño **P**, si la máquina efectivamente mejora su desempeño **P** en la tarea **T** a partir de la experiencia **E**.

Aprendizaje

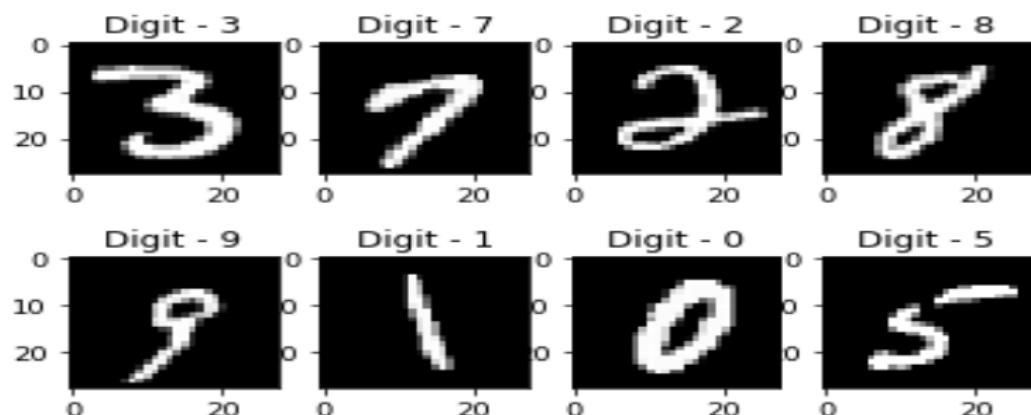
- Clase de tareas T
 - ▶ Reconocimiento de patrones
 - ▶ Creación de relaciones
 - ▶ Toma de decisiones
- La medida P y el tipo de experiencias E dividen los algoritmos en clases

Ejemplo: Reconocer dígitos escritos a mano



Ejemplo: Reconocer dígitos escritos a mano

- Tarea T: clasificar dígitos a partir de imágenes
- Medida de desempeño P: porcentaje de dígitos clasificados correctamente
- Experiencia de entrenamiento E: data base de dígitos con clasificaciones anotadas, MNIST



Posición en las ciencias de datos

Aprendizaje Automático

El Aprendizaje Automático es una derivación natural de la intersección entre la Estadística y la Ciencia de la Computación

- Inferencia estadística → modelar la población a partir de la muestra (datos).
- CS → algoritmos que realicen la tarea eficientemente, teniendo en cuenta la Arquitectura del sistema de computación y la Escalabilidad y Tratabilidad del problema.

Posición en las ciencias de datos

- **Inteligencia Artificial** Desarrollo de agentes autónomos que presentan comportamiento que describimos como inteligente.
- **Ciencias cognitivas** Estudio del aprendizaje en animales y humanos → Neurociencias.
- **Ciencia de Datos** Marco unificado para la recolección, estandarización, representación, almacenamiento, recuperación, procesamiento, clasificación, minería, extracción de información, visualización, etc...

Posición en las ciencias de datos



Cuan grande es Grande?

- **LARGE:** Grandes cantidades de datos → rasgo distintivo de Big Data
 - ▶ Almacenamiento paralelo distribuido en cluster → MapReduce (Hadoop)
 - ▶ Procesamiento paralelo → GPGPU
 - ▶ Ejemplo: LHC
- **WIDE:** Gran número de características (variables) → rasgo distintivo de ML → dimensionalidad

Tipos de Aprendizaje Automático

- Aprendizaje supervisado
- Aprendizaje no supervisado
- Aprendizaje semi-supervisado
- Aprendizaje activo
- Aprendizaje por refuerzo
- Aprendizaje evolutivo
- Aprendizaje de conceptos

Tipos de Aprendizaje Automático

Supervised Learning

- Labeled data
- Direct feedback
- Predict outcome/future

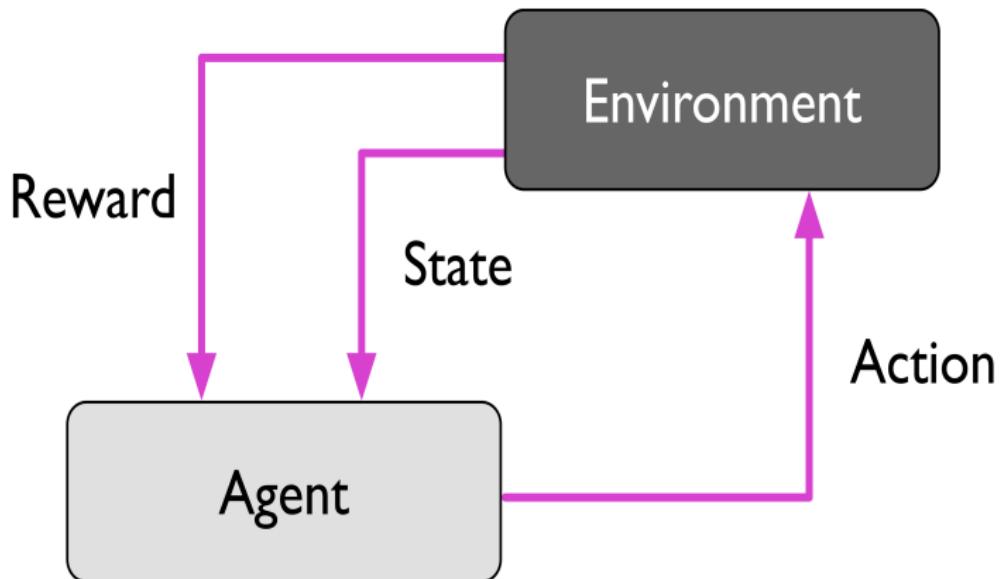
Unsupervised Learning

- No labels/targets
- No feedback
- Find hidden structure in data

Reinforcement Learning

- Decision process
- Reward system
- Learn series of actions

Tipos de Aprendizaje Automático



Tipos de Aprendizaje Automático

- **Aprendizaje evolutivo:** Inspirado en la evolución genética. Se generan variaciones llamadas mutaciones (mutations) del algoritmo que compiten entre sí midiendo su desempeño (fitness) en la tarea T bajo una misma experiencia E (environment).
- **Aprendizaje por refuerzo:** Se aplican en circunstancias donde las acciones tienen resultado indirecto: Jugar al ajedrez, videojuegos, conducir un vehículo (Planning).

Tipos de Aprendizaje Automático

- **Aprendizaje no supervisado** Detección automática de patrones inmersos en los datos y relaciones no conocidas entre ellos para transformarlas en una estructura comprensible → Data mining, Knowledge Discovery
 - ▶ Clustering
 - ▶ Reglas de Asociación

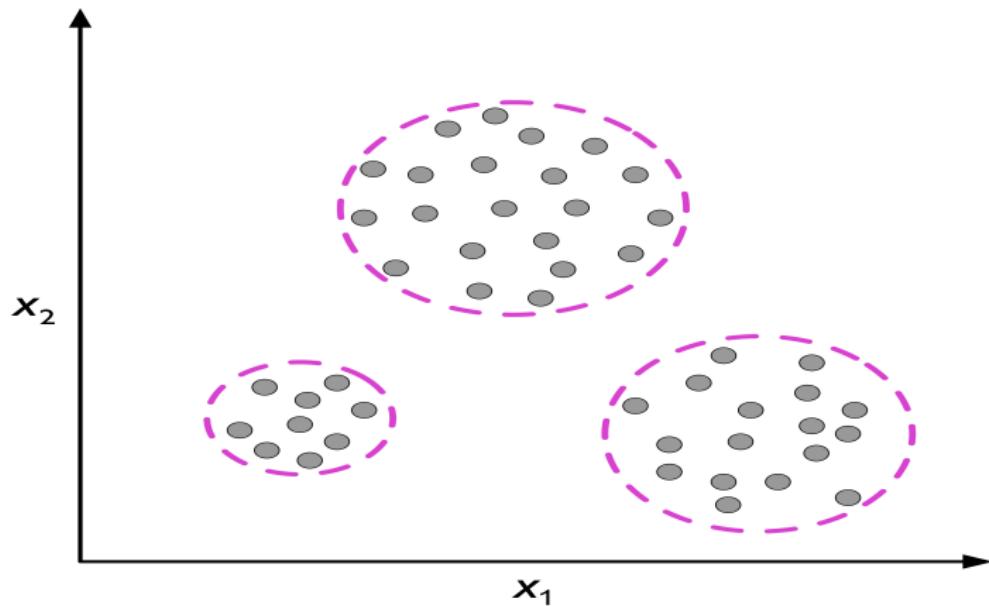
Tipos de Aprendizaje Automático

- **Aprendizaje supervisado** A partir de un conjunto de ejemplos de entrenamiento, etiquetados por expertos con la clase a la que pertenecen, los algoritmos deben generalizar para asignar correctamente la clase de pertenencia a toda entrada de la misma naturaleza.
 - ▶ Clasificación (clase con valores discretos)
 - ▶ Regresión (clase con valores continuos)

Aprendizaje no supervisado

- Estos algoritmos encuentran relaciones significativas entre las diferentes características de los datos usando:
 - ▶ semejanza
 - ▶ co-varianza (correlación) o independencia
 - ▶ probabilidad condicional
- procesando los datos para que un experto en el dominio los interprete proveyendo:
 - ▶ síntesis y sistematización de información
 - ▶ visualización
- y teniendo en cuenta criterios de fiabilidad:
 - ▶ estadísticamente significativo
 - ▶ reproducibilidad
 - ▶ interpretabilidad

Tipos de Aprendizaje Automático



Aprendizaje no supervisado

Ejemplos de detección de patrones:

- detección de anomalías en tráfico de red
- detección de bots
- segmentación de clientes
- identificación de gustos parecidos
- asociación de ítems en carritos de la compra
- descubrimiento de tendencias (trending topics)
- descubrimiento de productos equivalentes
- descubrimiento de keywords en colecciones documentales

Aprendizaje supervisado

Diseño de un clasificador automático (concepción teórica)

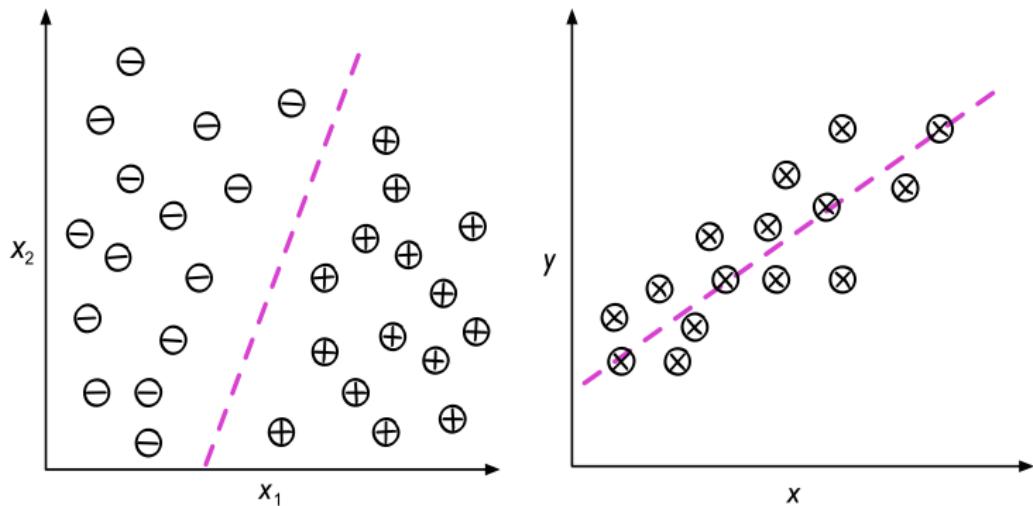
- Elección de la experiencia de aprendizaje **E** para la tarea **T**
 $\{x_{ij}, j = 1, \dots, M\}$ con $i = 1, \dots, N$, con N numero de ejemplos y M la dimensionalidad. El experto de dominio asigna la etiqueta y_i al elemento x_{ij} , que puede ser una de K clases.
- Elección de la función objetivo **F** (lo que se aprende)
 $y_i = F(x_{i1}, \dots, x_{iM})$
- Elección de la representación de la función objetivo **F**
- Elección del algoritmo que aproxima la función objetivo **F** en base a la medida de desempeño **P**

Aprendizaje supervisado

Implementación de un clasificador automático (realización práctica)

- a partir de ejemplos ← analizados por expertos
- se infiere como la información dada se asocia a la información deseada ← como las características determinan la clase
 - ▶ se supone que un cierto modelo es apropiado para capturar esta asociación (sesgo inductivo)
 - ▶ se ajusta el modelo con los datos, para que pueda predecir específicamente los ejemplos
 - ▶ se generaliza el modelo para que pueda predecir en nuevas instancias la clase adecuada
- este procedimiento dá por resultado un clasificador y se usa para asignar información (clase, valor, probabilidad) a nuevos ejemplos no analizados por expertos

Tipos de Aprendizaje Automático



Aprendizaje supervisado

Ejemplos de aprendizaje supervisado

- diagnóstico médico, acciones preventivas
- solicitudes de tarjetas de crédito
- cálculo de la cuota de un seguro (scoring)
- detección de fraude en transacciones electrónicas
- identificación de spam
- recomendación: artículos, películas, libros, música, productos
- inversiones financieras
- reconocimiento de lenguaje escrito y hablado
- etiquetado de imágenes
- detección de eventos en el LHC (Large Hadron Collider) del CERN

Reconocimiento de patrones

Construir una máquina capaz de reconocer patrones que mejore con la experiencias recibidas:

- Reconocimiento de frases habladas
- Identificación de huellas digitales
- Reconocimiento óptico de escritura
- Identificación de secuencias de ADN

Reconocimiento de Meros y Salmones

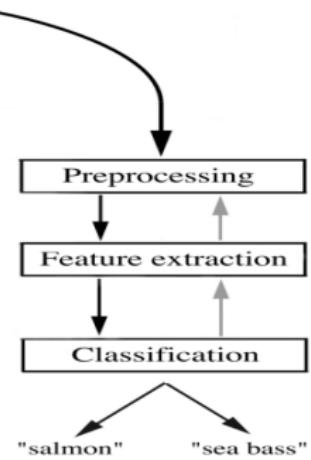
Ejemplo

Separar el pescado entrante en una cinta de transporte, de acuerdo con la especie, usando sensores ópticos

Especies:

Mero y Salmón.

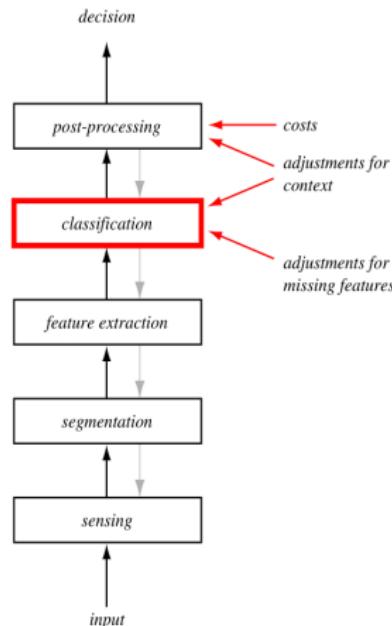
Reconocimiento de Meros y Salmones



Esquema general

- Montar una cámara
- Tomar algunas imágenes, procesarlas y etiquetar los datos
- Decidir que características se miden
- Diseñar un clasificador
- Ajustar resultados por costos y contexto
- Generar decisión

Reconocimiento de Meros y Salmones



Módulo de Procesamiento inicial

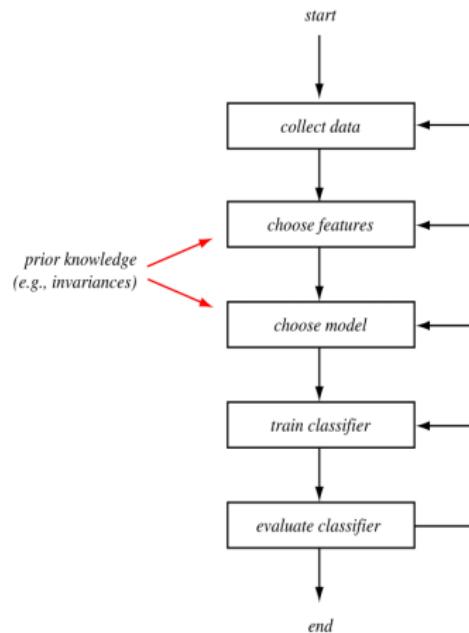
Los primeros dos puntos son el procesamiento inicial,

- Montaje de la cámara y calibración de la toma de imágenes.
- Ya obtenida la imagen, usar segmentación para aislar los peces uno de otro y con respecto al fondo
- Enviar información de cada pez al algoritmo extractor de características del tipo de
 - ▶ Largo
 - ▶ Ancho
 - ▶ Tamaño escamas
 - ▶ Tono

Módulo de clasificación

- Decisión y extracción de características
 - ▶ Características discriminativas
 - ▶ Características invariantes con respecto a la traslación, rotación y escala.
- Clasificación
 - ▶ Diseño de la regla que usa el vector de características provisto para asignar el objeto a una categoría
- Post Procesamiento
 - ▶ Explotar la información de contexto del objeto para mejorar el desempeño

Reconocimiento de Meros y Salmones



Clasificación

- Decidir el conjunto de características a usar
 - ▶ Conjunto suficiente
 - ▶ Conjunto mínimo
- Seleccionar el largo del pez como una posible característica para discriminación
- Clasificación lineal, $\text{largo}(x) < l^* \rightarrow x = \text{salmon}$

Clasificación

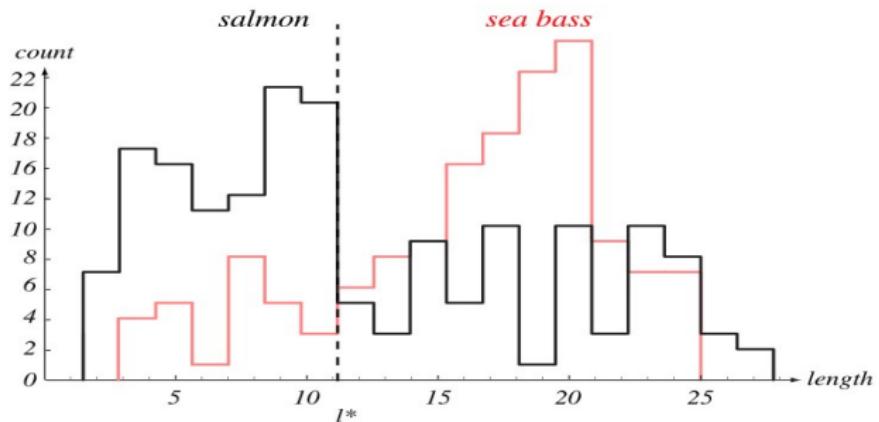


FIGURE 1.2. Histograms for the length feature for the two categories. No single threshold value of the length will serve to unambiguously discriminate between the two categories; using length alone, we will have some errors. The value marked l^* will lead to the smallest number of errors, on average. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Clasificación

- El largo es una característica muy pobre!
- Otras características
 - ▶ Largo
 - ▶ Tono (color)
 - ▶ Tamaño
 - ▶ Número y forma de las aletas
 - ▶ Posición de la boca, etc
 - ▶ Tono de las escamas posible característica.
- Clasificación lineal, $\text{Tono}(x) < x^*$ $\longrightarrow x = \text{salmón}$

Clasificación

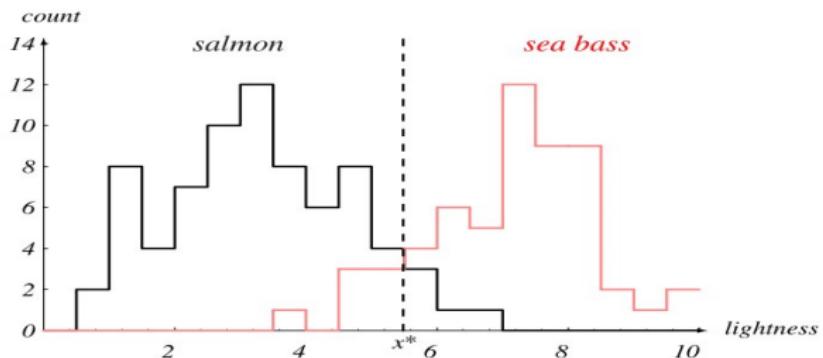


FIGURE 1.3. Histograms for the lightness feature for the two categories. No single threshold value x^* (decision boundary) will serve to unambiguously discriminate between the two categories; using lightness alone, we will have some errors. The value x^* marked will lead to the smallest number of errors, on average. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Umbral de decisión x^* y Costo de clasificación

- Muy importante no etiquetar meros por salmones.
- Mover nuestro borde de decisión x^* hacia valores menores del Tono para minimizar el costo (reducir el numero de Meros que son clasificados como Salmones !)

Clasificación lineal

- Todavía hay salmones mal clasificados.
- Puedo utilizar un par de variables, Tono y Ancho de escamas. $X = [x_1, x_2]$.
- La decisión lineal es clasifico en Salmón si $x_2 < ax_1 + b$

Clasificación

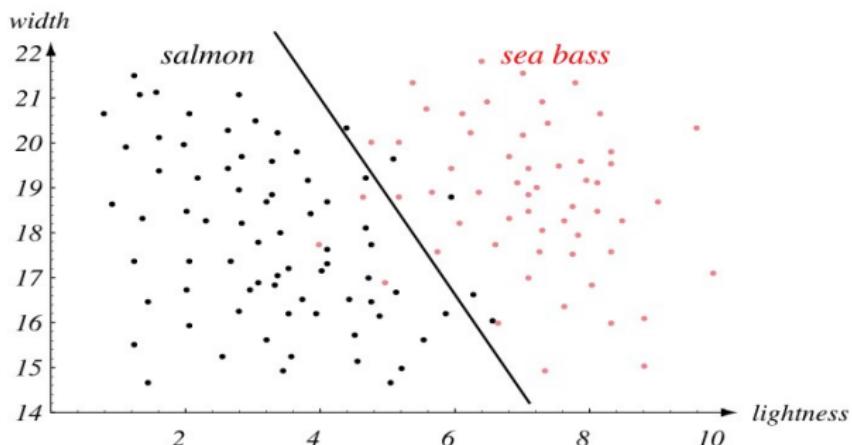


FIGURE 1.4. The two features of lightness and width for sea bass and salmon. The dark line could serve as a decision boundary of our classifier. Overall classification error on the data shown is lower than if we use only one feature as in Fig. 1.3, but there will still be some errors. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Otras opciones

- Podríamos agregar otras características que sean no correlacionadas con las que ya tenemos.
- Hay que tomar precauciones para no reducir el desempeño agregando características ruidosas.
- Idealmente, el mejor borde de decisión debería ser el que provee optimo desempeño como en la figura siguiente :

Clasificación

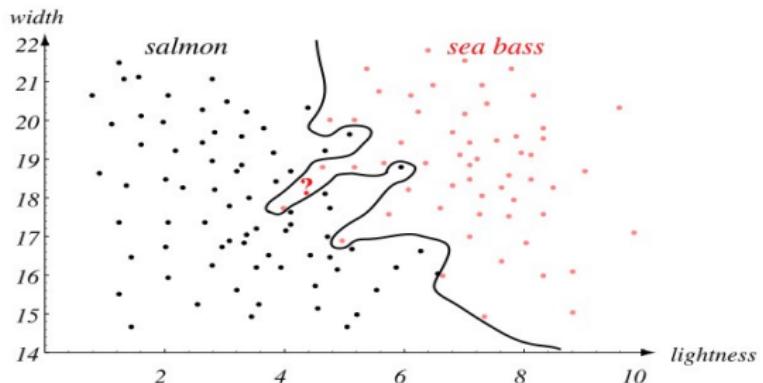


FIGURE 1.5. Overly complex models for the fish will lead to decision boundaries that are complicated. While such a decision may lead to perfect classification of our training samples, it would lead to poor performance on future patterns. The novel test point marked ? is evidently most likely a salmon, whereas the complex decision boundary shown leads it to be classified as a sea bass. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Otras opciones

- Sin embargo, nuestra satisfacción es prematura porque el objetivo central es designar un clasificador que clasifique correctamente un nuevo elemento
- Se debe poder generalizar

Tipo de dato condiciona el clasificador

- Las variables de este ejemplo son contínuas.
- Si las variables son categóricas, otras estrategias son mas eficientes.
- Puede considerarse al clasificador como un descriptor de un concepto.
- Estos descriptores son determinísticos, no consideran una distribución de probabilidad en los datos, por lo cual son muy sensibles al ruido.