Teoría de Aprendizaje

Lección 14

Dr. Pablo Alvarado Moya

CE5506 Introducción al reconocimiento de patrones
Área de Ingeniería en Computadores
Tecnológico de Costa Rica

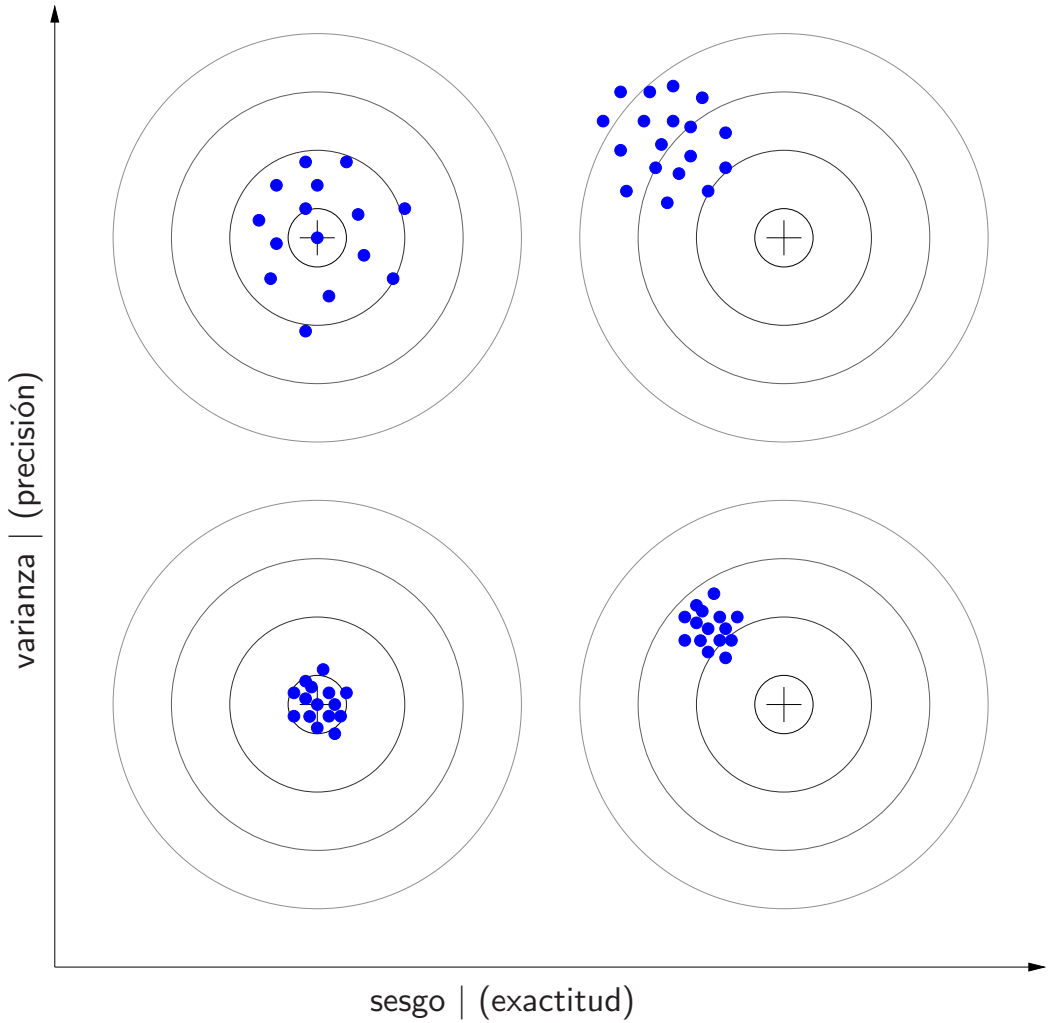
II Semestre, 2019

Contenido

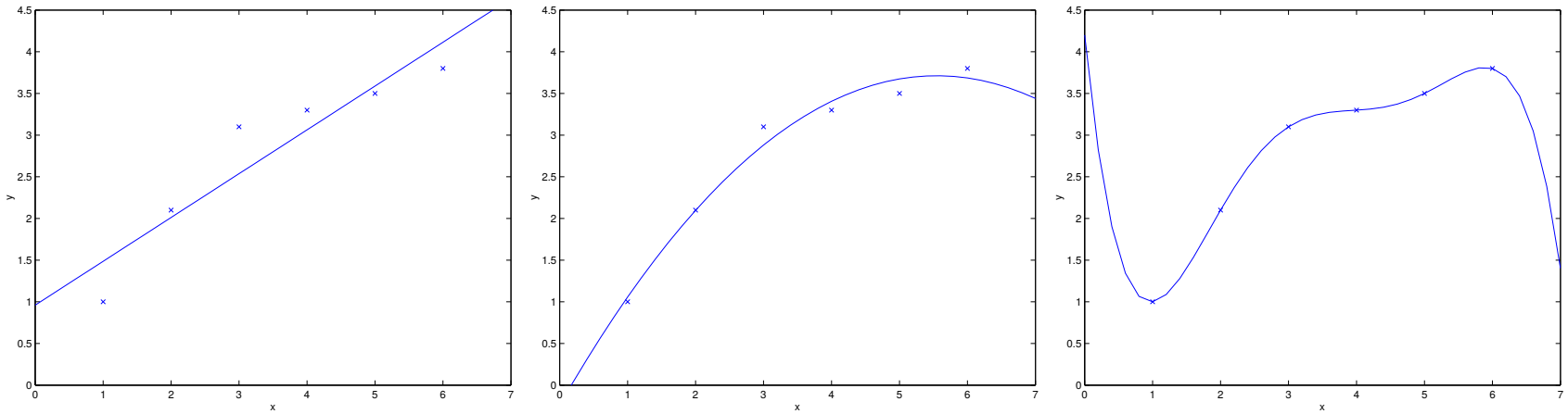
Introducción

- Ya hemos revisado gran variedad de algoritmos supervisados
- Estos son "solo" un conjunto de herramientas
- Ahora necesitamos aprender cómo usar esas herramientas
- Revisaremos propiedades de algoritmos de aprendizaje que permitan decidir cuándo usar cuál

Sesgo y varianza



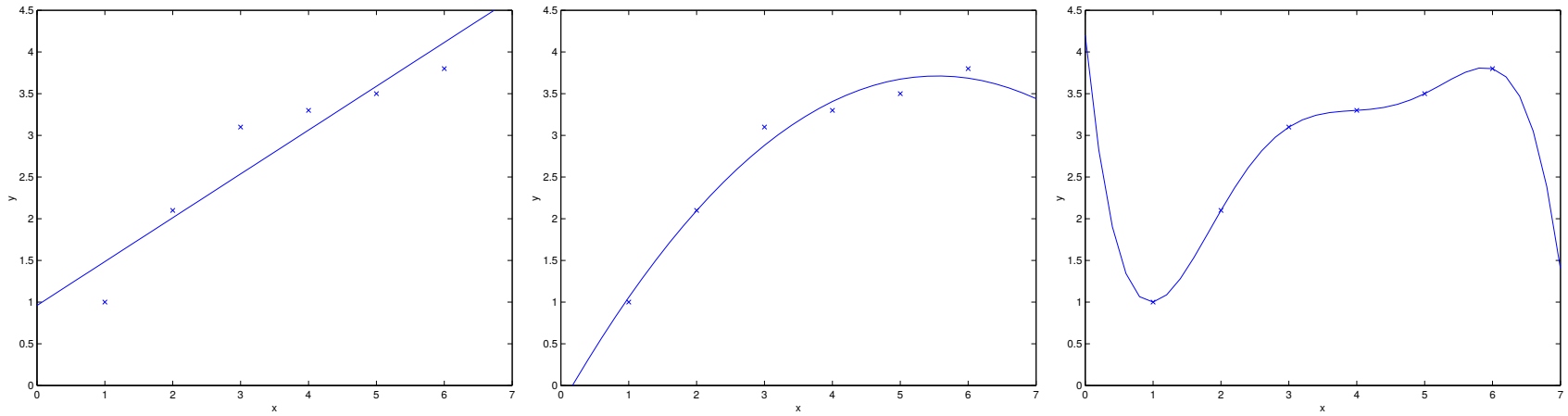
Sesgo y varianza



(a) (b) (c)

- El modelo en (a) es **simple**, en el sentido de que solo puede predecir una estructura de línea. Tiene pocos parámetros.
- Decimos que hay **alto sesgo** (*high bias*), porque estamos "sesgando" predicciones a un modelo supuesto a-priori
- Algo sesgo lleva a **sub-ajuste** (*underfitting*) a los datos.

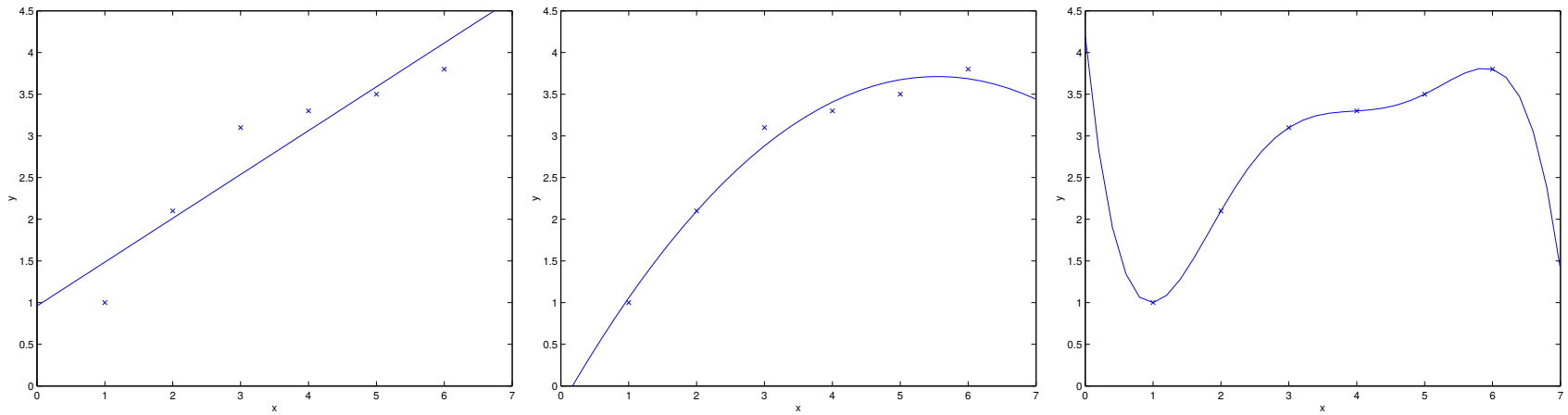
Sesgo y varianza



(a) (b) (c)

- El alto sesgo implica usualmente que habrá alto error de generalización, aún si entrenamos con muchos datos

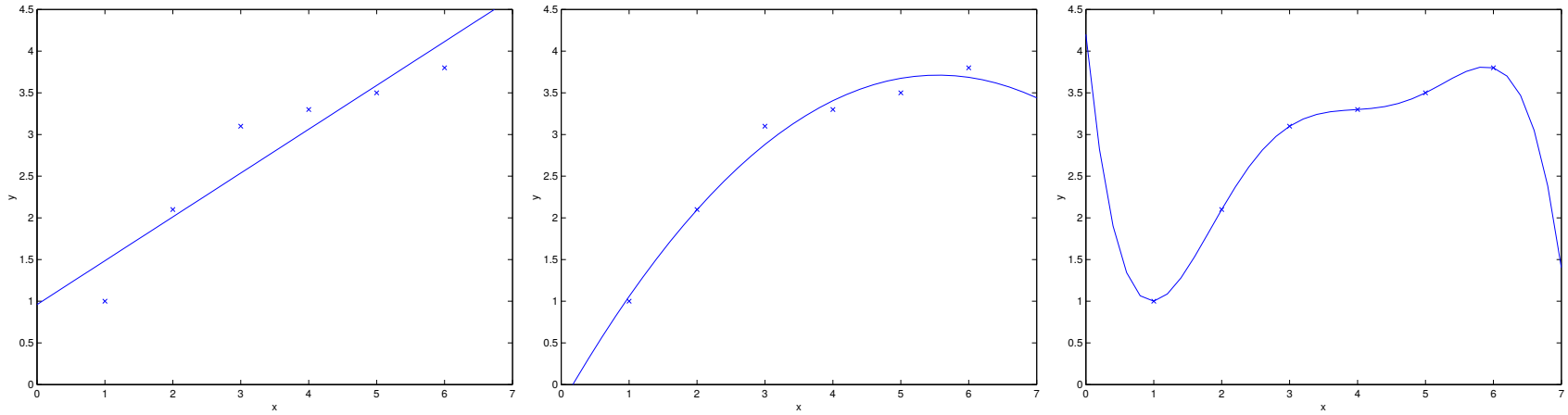
Sesgo y varianza



(a) (b) (c)

- El modelo en (c) es **complejo**. Tiene más parámetros.
- Complejidad permite ajustar más tipos de curvas.
- Decimos que tiene **alta varianza** (*high variance*) por la mayor variabilidad alcanzable por modelo
- Modelo más complejo se **sobre-ajusta** a datos (*overfitting*)

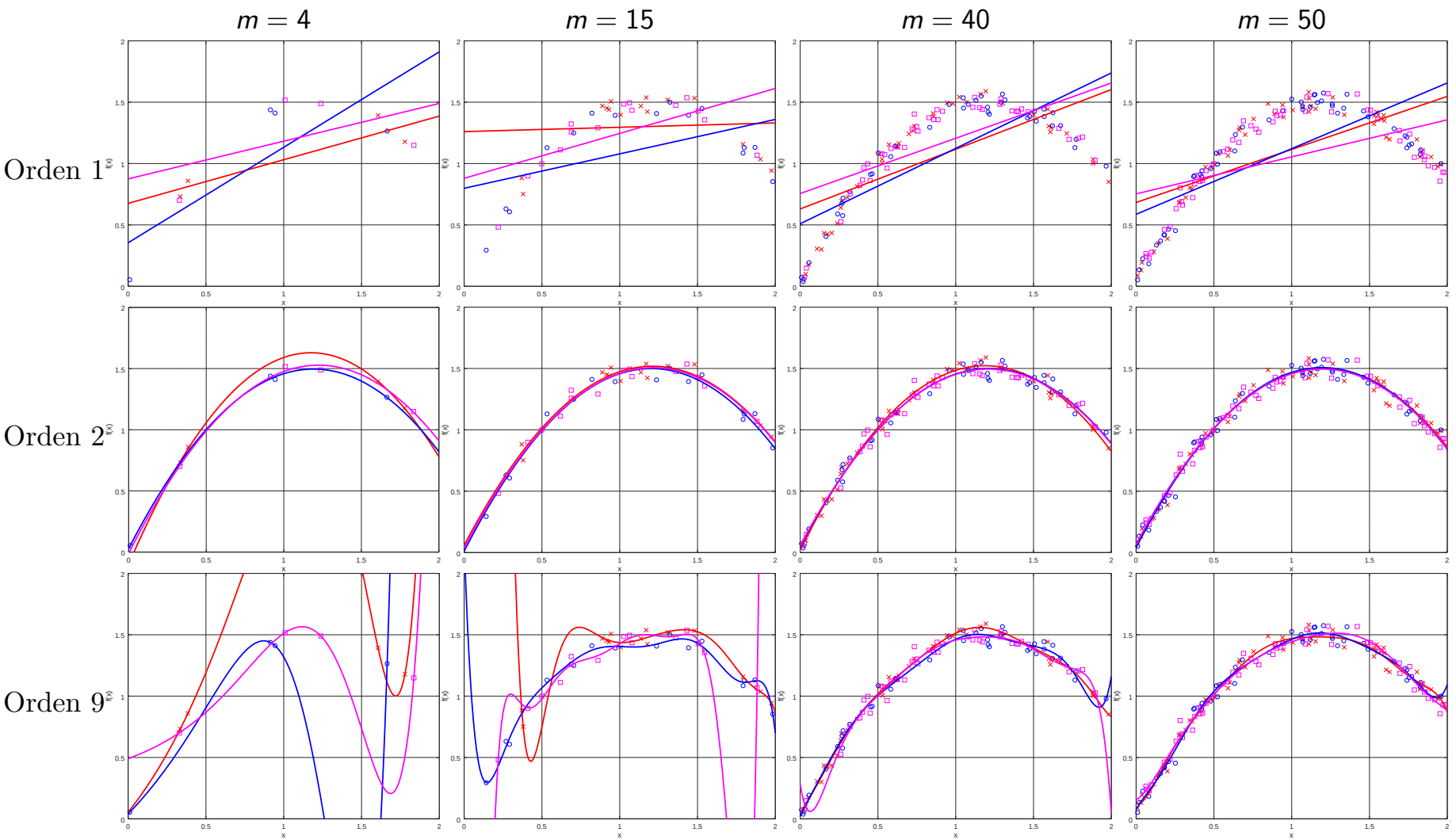
Sesgo y varianza



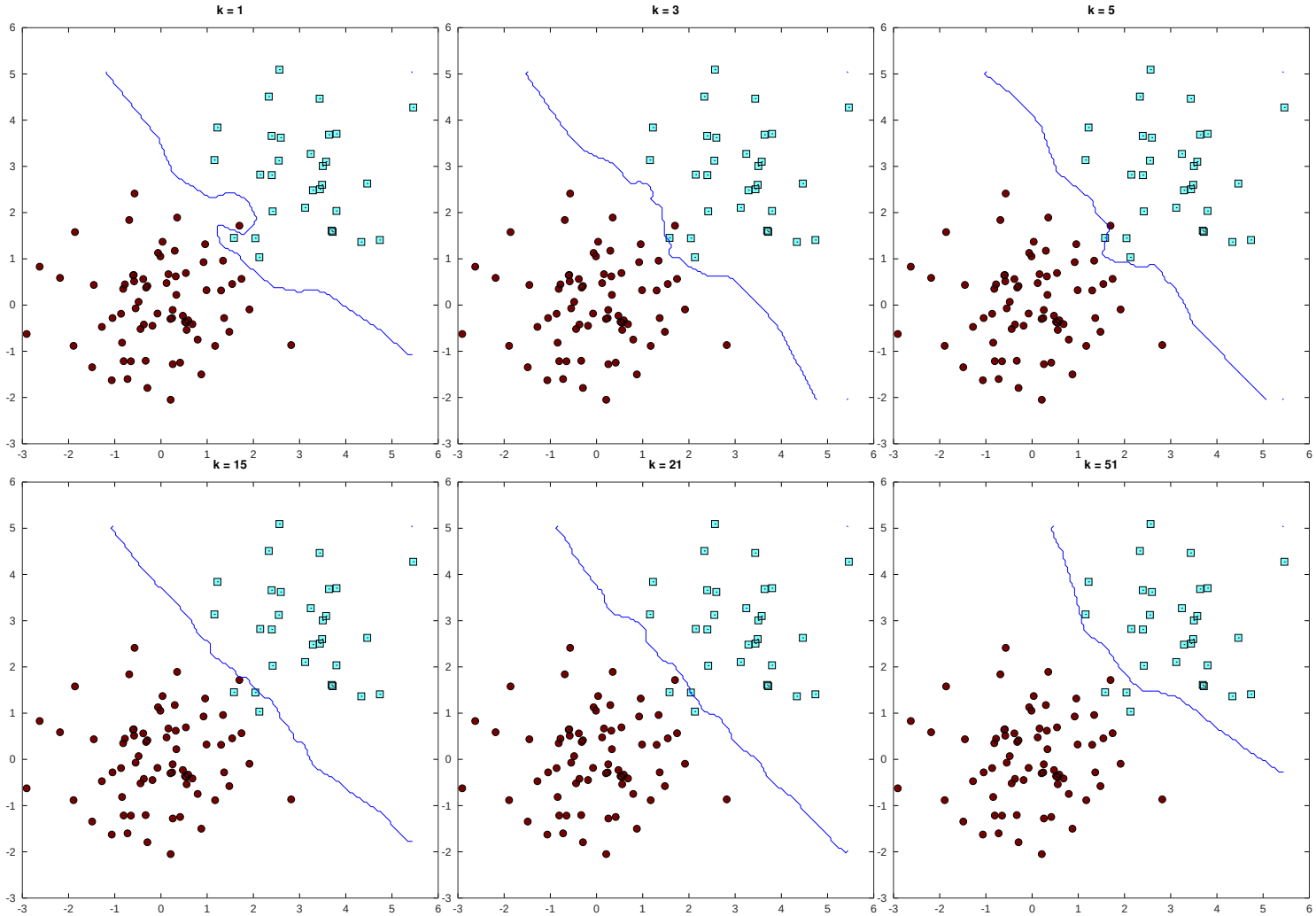
(a) (b) (c)

- Principio de parsimonia (Ockham's razor) *En igualdad de condiciones, la explicación más sencilla suele ser la más probable*

Ejemplo de regresión lineal con hipótesis $h_{\underline{\theta}} = \underline{\theta}^T \phi(\underline{\mathbf{x}})$



Ejemplo de clasificación con kNN



Clasificación lineal

- Supongamos un clasificador lineal con

$$h_{\underline{\theta}} = g(\underline{\theta}^T \underline{\mathbf{x}})$$
$$g(z) = 1\{z \geq 0\}$$

- Volvemos a salidas $y \in \{0,1\}$ y entradas con $x_0 = 1$
- Conjunto de entrenamiento es $\mathcal{S} = \{(\underline{\mathbf{x}}^{(i)}, y^{(i)}) \mid i = 1, \ldots m\}$
- $(\underline{\mathbf{x}}^{(i)}, y^{(i)}) \sim_{i.i.d} \mathcal{D}$
- Para entender mejor sesgo y varianza vamos a usar un modelo simplificado de aprendizaje

Riesgo

- El error empírico de aprendizaje (o **riesgo** empírico, o error de entrenamiento) lo definimos como

$$\hat{\varepsilon}(h_{\underline{\theta}}) = \hat{\varepsilon}_{\mathcal{S}}(h_{\underline{\theta}}) = \frac{1}{m}\sum_{i=1}^{m} 1\left\{h_{\underline{\theta}}(\underline{\mathbf{x}}^{(i)}) \neq y^{(i)}\right\}$$

  (fracción de datos de entrenamiento mal clasificados)
- Buscamos minimización del riesgo empírico (*empirical risk minimization (ERM)*)

$$\hat{\underline{\theta}} = \arg\min_{\underline{\theta}} \hat{\varepsilon}(h_{\underline{\theta}})$$

- Probaremos propiedades sobre este problema de optimización
- SVM y regresión logística *aproximan* al ERM
- ERM es un problema de optimización no convexo, NP-hard

Replanteamiento de ERM (1)

- Cambiemos estrategia de ERM: en vez de buscar **parámetros**, busquemos una **función**
- Definamos la clase de hipótesis $\mathcal{H}$ como un conjunto de clasificadores.
- Por ejemplo la clase de clasificadores lineales es:

$$\mathcal{H} = \{h_{\underline{\theta}} \mid h_{\underline{\theta}}(\underline{\mathbf{x}}) = 1\left\{\underline{\theta}^T \underline{\mathbf{x}} \geq 0\right\}; \underline{\theta} \in \mathbb{R}^{n+1}\}$$

con $h_{\underline{\theta}} : \mathbb{X} \rightarrow \{0,1\}$

con $\mathbb{X}$ el dominio de entrada tal que $\underline{\mathbf{x}} \in \mathbb{X} \subseteq \mathbb{R}^{n+1}$

Replanteamiento de ERM (2)

- Redefinamos ERM como la selección de la *mejor* **función**:

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\varepsilon}(h)$$

- Ventaja de esta representación es que $\mathcal{H}$ puede contener cualquier tipo de clasificador, incluso combinar tipos
- Aquí usaremos clasificación binaria, pero conceptos fácilmente generalizables a regresión y clasificación multi-clase

Generalización

- El **riesgo** no es exactamente lo que nos interesa (solo considera conjunto de entrenamiento)
- Preferimos evaluar la **generalización**: ¿qué tan bien predecimos con entradas que no hemos visto antes?
- Error de generalización:

$$\varepsilon(h) = P_{(\underline{\mathbf{x}},y)\sim\mathcal{D}}(h(\underline{\mathbf{x}}) \neq y)$$

  es decir, la probabilidad de que si tomamos cualquier par $(\underline{\mathbf{x}},y)$ de la distribución $\mathcal{D}$, la hipótesis h lo clasifique mal.
- Note que estamos asumiento que los datos de entrenamiento y de prueba los tomamos de la misma distribución $\mathcal{D}$

  (esta es una de las suposiciones PAC (*probably approximately correct*), que es un marco teórico desarrollado en los 80 para analizar aprendizaje automático)

Objetivos

- Queremos contestar varias preguntas:
  - ¿Qué relación hay entre el riesgo empírico y el error de generalización?
    (o ¿de qué nos sirve medir error con datos de entrenamiento?)
  - ¿Existen condiciones bajo las cuales podemos demostrar que un algoritmo de aprendizaje funcionará bien?
- Para poder formalizar relaciones entre sesgo, varianza, riesgo empírico y error de generalización necesitamos definir algunos conceptos

Cota de unión

- La cota de unión (*union bound*) establece que la probabilidad de que cualquiera de k eventos suceda es a lo sumo la suma de las probabilidades de ocurrencia de los k eventos por separado
- Sean k eventos distintos $A_1, A_2, \ldots, A_k$, entonces:

$$P(A_1 \cup \ldots \cup A_k) \leq P(A_1) + \ldots + P(A_k)$$

- Esto es uno de los axiomas de la probabilidad, por lo que, aunque intuitivo, no se suele demostrar.

Desigualdad de Hoeffding

- Sean $Z_1, \ldots, Z_m$ m variables aleatorias i. i. d. tomadas de una distribución de Bernoulli(ϕ) (esto es $P(Z_i = 1) = \phi$, $P(Z_i = 0) = 1 - \phi$).
- Sea

$$\hat{\phi} = \frac{1}{m} \sum_{i=1}^{m} Z_i$$

  la media de esas variables y sea $\gamma > 0$ cualquier constante.
- La desigualdad de Hoeffding establece:

$$P(|\phi - \hat{\phi}| > \gamma) \leq 2 \exp(-2\gamma^2 m)$$

- Nótese que m afecta decrecimiento de $P(|\phi - \hat{\phi}| > \gamma)$: mientras mayor m, más rápido decrece

Caso de $\mathcal{H}$ finito

Minimización de riesgo empírico

- Caso de clases de hipótesis **finitas**
- $\mathcal{H} = \{h_1, h_2, \ldots, h_k\}$, k hipótesis
- ERM toma conjunto de entrenamiento, y toma de $\mathcal{H}$ la h_i con menor error:

$$\hat{h} = \arg \min_{h_i \in \mathcal{H}} \hat{\varepsilon}(h_i)$$

- Estrategia:
  1. Mostrar que $\hat{\varepsilon}(h)$ aproxima a $\varepsilon(h)$ para todo $h \in \mathcal{H}$
  2. Mostrar que hay cota superior para error de generalización $\varepsilon(\hat{h})$

Caso para una hipótesis

- Tomemos una hipótesis cualquiera $h_j \in \mathcal{H}$ fija
- Definamos variable aleatoria Bernoulli Z así:
  1. Generemos $(\underline{\mathbf{x}},y) \sim_{iid} \mathcal{D}$ (mismo proceso del entrenamiento)
  2. $Z = 1\{h_j(\underline{\mathbf{x}}) \neq y\} \in \{0,1\}$
  3. De forma similar definamos $Z_i = 1\left\{h_j(\underline{\mathbf{x}}^{(i)}) \neq y^{(i)}\right\} \in \{0,1\}$
- Como Z y Z_i se generaron de muestras i. i. d. de $\mathcal{D}$, entonces ellas mismas son ambas i. i. d. y siguen la misma distribución de Bernoulli
- La probabilidad de clasificación errónea de una muestra aleatoria $(\underline{\mathbf{x}},y)$ cualquiera la hemos llamado $\varepsilon(h)$, y equivale (por ser Z Bernoulli) al valor esperado de Z (o Z_i):

$$P(Z_i = 1) = P(Z = 1) = \varepsilon(h_j)$$

Cota de error de entrenamiento para una hipótesis

- El error de entrenamiento se puede reescribir como:

$$\hat{\varepsilon}(h_j) = \frac{1}{m}\sum_{i=1}^{m} Z_i = \frac{1}{m}\sum_{i=1}^{m} 1\left\{h_j(\underline{\mathbf{x}}^{(i)}) \neq y^{(i)}\right\}$$

- $\hat{\varepsilon}(h_j)$ es la media de m variables aleatorias Z_i tomadas i. i. d. de una distribución de Bernoulli, cada una con media $\varepsilon(h_j)$
- Si aplicamos la desigualdad de Hoeffding obtenemos:

$$P(|\varepsilon(h_j) - \hat{\varepsilon}(h_j)| > \gamma) \leq 2\exp(-2\gamma^2 m)$$

- Esto quiere decir que la probabilidad de que los errores de entrenamiento y generalización estén cerca aumenta conforme m crece.

Convergencia uniforme (1)

- ¿Se cumplirá lo anterior para **todas** las hipótesis $h \in \mathcal{H}$?
- Sea A_j un evento tal que $|\varepsilon(h_j) - \hat{\varepsilon}(h_j)| > \gamma$
- Ya probamos que $P(A_j) \leq 2\exp(-2\gamma^2 m)$
- Con la cota de unión sabemos que

$$\begin{aligned} P(\exists h_j \in \mathcal{H} \cdot |\varepsilon(h_j) - \hat{\varepsilon}(h_j)| > \gamma) &= P(A_1 \cup \ldots \cup A_k) \\ &\leq \sum_{i=1}^{k} P(A_i) \leq \sum_{i=1}^{k} 2\exp(-2\gamma^2 m) \\ &= 2k\exp(-2\gamma^2 m) \end{aligned}$$

Convergencia uniforme (2)

- Restando de 1 a ambos lados tenemos

$$P(\nexists h \in \mathcal{H} \cdot |\varepsilon(h) - \hat{\varepsilon}(h)| > \gamma) = P(\forall h \in \mathcal{H} \cdot |\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma)$$
$$\geq 1 - 2k \exp(-2\gamma^2 m)$$

- A esto se le llama **convergencia uniforme**
- La convergencia uniforme nos da la cota de probabilidad para todas las k hipótesis $h \in \mathcal{H}$ de que el error de entrenamiento se encuentre a distancia γ del error de generalización.
- Tres términos están relacionados por la convergencia uniforme: probabilidad de error, γ y m
- Podemos buscar cada uno de ellos en términos de los otros

Cota de complejidad muestral

- Dados γ y un $\delta > 0$, ¿qué tan grande debe ser m para garantizar con probabilidad $1 - \delta$ que el error de entrenamiento estará a distancia γ del error de generalización?
- Haciendo $\delta \geq 2k \exp(-2\gamma^2 m)$ se despeja m:

$$m \geq \frac{1}{2\gamma^2} \ln \frac{2k}{\delta}$$

- Con ese valor de m, con probabilidad al menos $1 - \delta$, tenemos que $|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma$ para todo $h \in \mathcal{H}$
- La **complejidad muestral** es el tamaño del conjunto de entrenamiento requerido por un método/algoritmo para alcanzar un cierto nivel de desempeño.
- Nótese que en la cota anterior, el número de muestras requerido crece *solo* con el logaritmo de k.

Cota de error

Error bound

- La cota de error especifica la divergencia entre errores de generalización y de entrenamiento en términos de m y δ.
- Con probabilidad $1 - \delta$ se tiene para toda hipótesis $h \in \mathcal{H}$

$$|\hat{\varepsilon}(h) - \varepsilon(h)| \leq \sqrt{\frac{1}{2m} \ln \frac{2k}{\delta}}$$

Error de generalización de h con menor $\hat{\varepsilon}(h)$

- Asumamos que tenemos convergencia uniforme, es decir, $\forall h \in \mathcal{H}, |\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma$.
  ¿Qué podemos probar acerca de la generalización de la hipótesis seleccionada con $\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\varepsilon}(h)$ (ERM)?
- Sea $h^* = \arg\min_{h \in \mathcal{H}} \varepsilon(h)$ la mejor hipótesis posible en $\mathcal{H}$
- ¿Qué relación hay entre $\hat{h}$ y h^*? Se cumple:

$$\begin{aligned}
\varepsilon(\hat{h}) &\leq \hat{\varepsilon}(\hat{h}) + \gamma && \text{usando } |\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma \\
&\leq \hat{\varepsilon}(h^*) + \gamma && \text{usando } \hat{\varepsilon}(\hat{h}) \leq \hat{\varepsilon}(h), \forall h \Rightarrow \hat{\varepsilon}(\hat{h}) \leq \hat{\varepsilon}(h^*) \\
&\leq \varepsilon(h^*) + 2\gamma && \text{usando la primera expresión}
\end{aligned}$$

- $\Rightarrow$ Si hay convergencia uniforme, el error de generalización de $\hat{h}$ es a lo sumo 2γ peor que la mejor hipótesis en $\mathcal{H}$

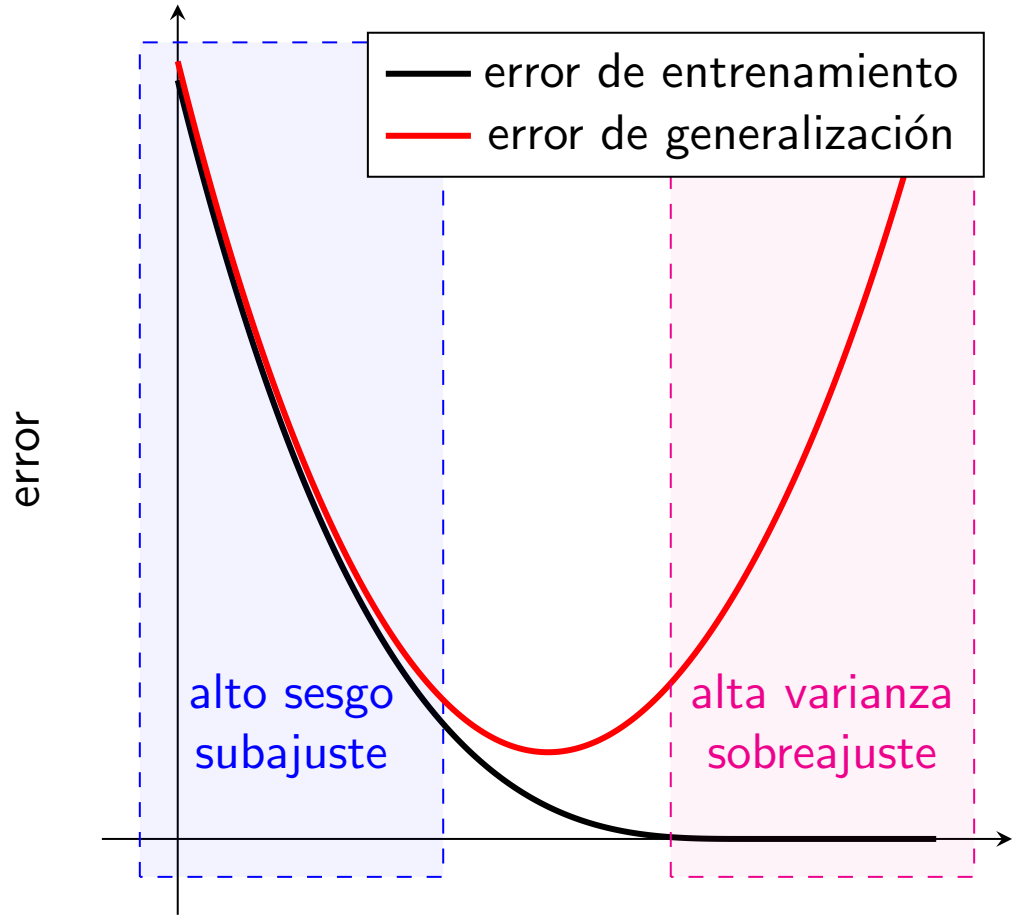
Teorema

- Sea $|\mathcal{H}| = k$; con m y δ fijos.
- Entonces, con $\varepsilon(\hat{h}) \leq \varepsilon(h^*) + 2\gamma$ y con probabilidad de al menos $1 - \delta$ tenemos

$$\varepsilon(\hat{h}) \leq \underbrace{\left(\min_{h \in \mathcal{H}} \varepsilon(h)\right)}_{\varepsilon(h^*)} + \underbrace{2\sqrt{\frac{1}{2m} \ln \frac{2k}{\delta}}}_{\gamma}$$

- Esto cuantifica el compromiso entre sesgo y varianza en selección de modelos
- Si nos pasamos a una clase de hipótesis mayor $\mathcal{H}' \supseteq \mathcal{H}$, entonces h^* solo puede bajar (baja el sesgo)
- Pero para $\mathcal{H}'$ el k es mayor y por tanto γ es mayor (sube la varianza)

Gráfico para m fijo



Corolario

- Sea $|\mathcal{H}| = k$ y sean δ y γ fijos.
- Para que se cumpla $\varepsilon(\hat{h}) \leq \text{mín}_{h \in \mathcal{H}} \varepsilon(h) + 2\gamma$ con probabilidad de al menos $1 - \delta$, es suficiente que

$$\begin{aligned} m &\geq \frac{1}{2\gamma^2} \ln \frac{2k}{\delta} \\ &= \mathcal{O}\left(\frac{1}{\gamma^2} \ln \frac{2k}{\delta}\right) \end{aligned}$$

El caso de $\mathcal{H}$ infinito (1)

- Probamos cotas interesantes para $|\mathcal{H}| = k$
- En todos los clasificadores que hemos visto tenemos infinitas posibilidades
- Lo que sigue **no** es el argumento formal, pero nos da la idea...
- Supongamos que $\mathcal{H}$ está parametrizada por d números reales
- Puesto que usamos representaciones de 64 bits de punto flotante por parámetro, en realidad tenemos $k = 2^{64d}$ posibles hipótesis

El caso de $\mathcal{H}$ infinito (2)

- Del corolario anterior

$$m \geq \mathcal{O}\left(\frac{1}{\gamma^2} \ln \frac{2k}{\delta}\right)$$

tenemos que, para garantizar con probabilidad de al menos $1-\delta$ que $\varepsilon(\hat{h}) \leq \varepsilon(h^*) + 2\gamma$, entonces es suficiente que

$$m \geq \mathcal{O}\left(\frac{1}{\gamma^2} \ln \frac{2^{64d}}{\delta}\right) = \mathcal{O}\left(\frac{d}{\gamma^2} \ln \frac{1}{\delta}\right) = \mathcal{O}_{\gamma,\delta}(d)$$

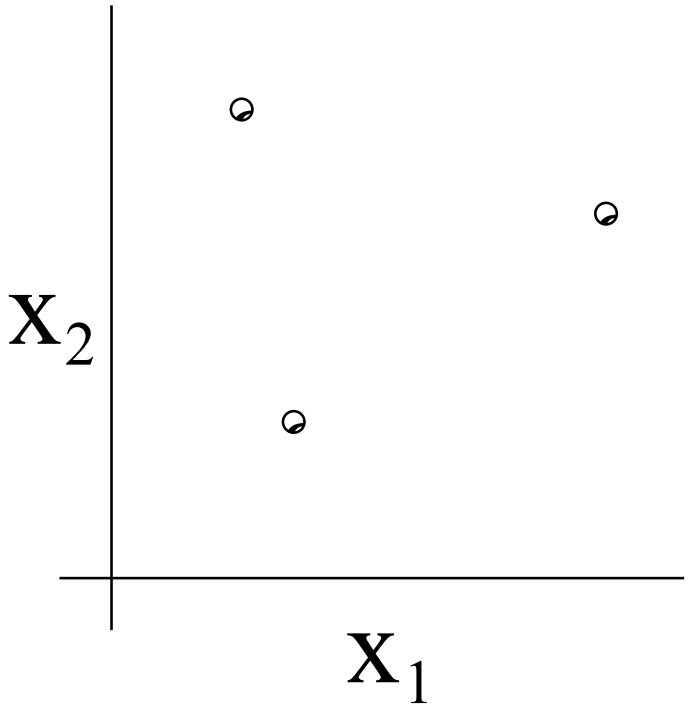
- $\mathcal{O}_{\gamma,\delta}$ denota que las constantes dependen de γ y δ
- La demostración formal es compleja, pero la intuición de que m crece linealmente con d es más o menos correcta

Separación

- Definición: Sea $\mathcal{S} = \{\underline{\mathbf{x}}^{(1)}, \underline{\mathbf{x}}^{(2)}, \ldots, \underline{\mathbf{x}}^{(d)}\}$ un conjunto de d puntos $\underline{\mathbf{x}}^{(i)} \in \mathbb{X}$ (ninguna relación con conjunto de entrenamiento)
- Decimos que $\mathcal{H}$ **separa** (*shatters*) a $\mathcal{S}$, si $\mathcal{H}$ puede encontrar cualquier etiquetación de $\mathcal{S}$, es decir, para cualquier conjunto de etiquetas $\{y^{(1)}, (2) \ldots, y^{(d)}\}$, existe $h \in \mathcal{H}$ tal que $h(\underline{\mathbf{x}}^{(i)}) = y^{(i)}$ para todo $i = 1, \ldots, d$

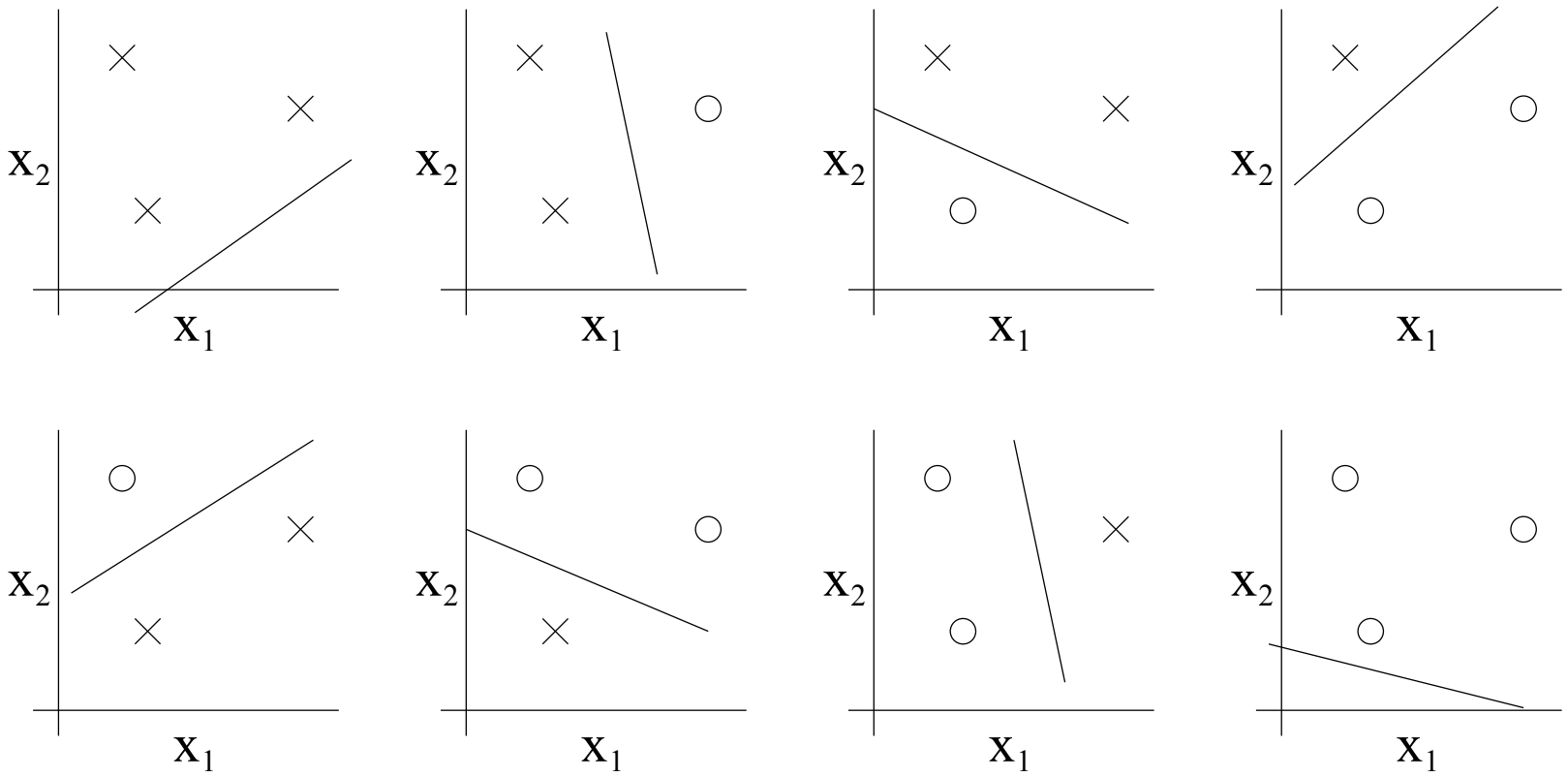
Ejemplo de separación

Dados los siguientes 3 puntos



¿Puede la clase $\mathcal{H}$ de clasificadores lineales $h(\underline{\mathbf{x}}) = 1\{\theta_0 + \theta_1 x_1 + \theta_2 x_2\}$ separarlos?

Ejemplo de separación

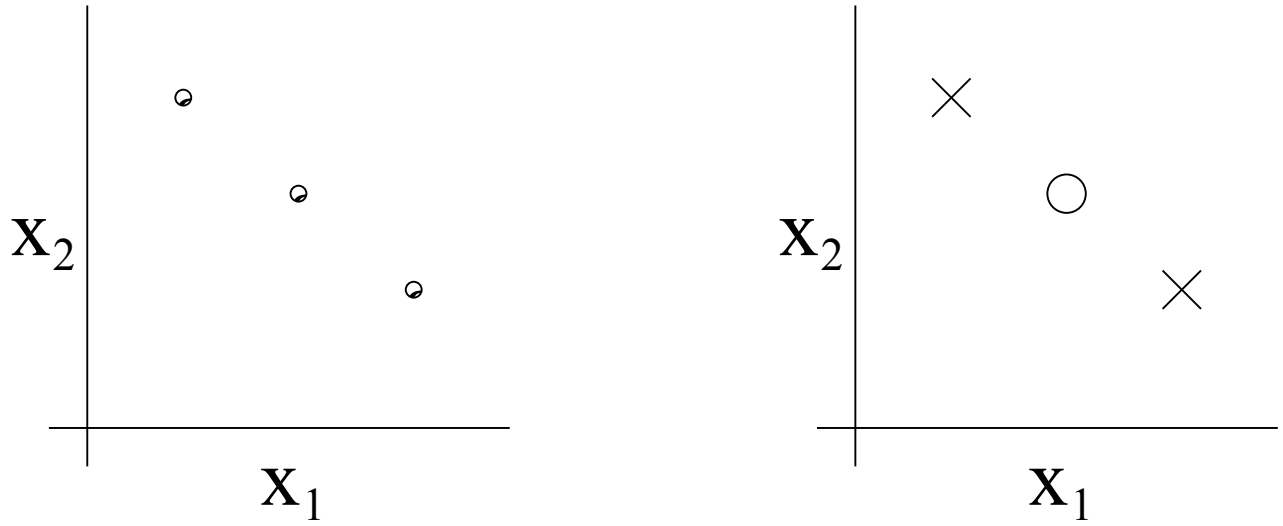


Dimensión Vapnik-Chervonenkis

- Dada una clase de hipótesis $\mathcal{H}$, definimos su **dimensión Vapnik-Chervonenkis** $\text{VC}(\mathcal{H})$ como el mayor tamaño de conjunto que es separable por $\mathcal{H}$
- Nóte que la clase de clasificadores lineales en 2D tiene $\text{VC}(\mathcal{H}) = 3$, pues podemos separar conjuntos de 2 o 3 puntos, pero no podemos separar un conjunto de 4 puntos.
- En n dimensiones, la clase de clasificadores lineales tiene $\text{VC}(\mathcal{H}) = n + 1$
- Si $\mathcal{H}$ puede separar conjuntos arbitrariamente grandes, entonces $\text{VC}(\mathcal{H}) = \infty$

Necesidad de un solo conjunto

- La dimensión VC es 3 aún si existen conjuntos de 3 puntos que $\mathcal{H}$ no puede separar:



Para que $\mathcal{H}$ sea de VC$(\mathcal{H}) = d$ solo tiene que existir **un** conjunto de d elementos que sea separable por $\mathcal{H}$

Teorema de Vapnik

- Este es quizá el teorema más importante de la teoría de aprendizaje automático
- Dado una clase $\mathcal{H}$ y sea $d = \text{VC}(\mathcal{H})$, con probabilidad al menos $1 - \delta$ para toda $h \in \mathcal{H}$

$$|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \mathcal{O}\left(\sqrt{\frac{d}{m}\ln\frac{m}{d} + \frac{1}{m}\ln\frac{1}{\delta}}\right)$$

- De este modo:

$$\varepsilon(\hat{h}) \leq \varepsilon(h^*) + \mathcal{O}\left(\sqrt{\frac{d}{m}\ln\frac{m}{d} + \frac{1}{m}\ln\frac{1}{\delta}}\right)$$

- En otras palabras, si una hipótesis tiene dimensión VC finita, entonces la convergencia uniforme ocurre cuando m se hace grande.
- Esto da una cota para $\varepsilon(h)$ en términos de $\varepsilon(h^*)$

Corolario

- Para garantizar que se cumple con probabilidad al menos $1-\delta$ que $|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma$ para todo $h \in \mathcal{H}$ (y entonces $\varepsilon(\hat{h}) \leq \varepsilon(h^*) + 2\gamma$) es suficiente que $m = \mathcal{O}_{\gamma,\delta}(d)$
- Para aprender bien usando $\mathcal{H}$, el número de datos de entrenamiento es de orden lineal con la dimensión VC de $\mathcal{H}$.
- Para la mayoría de clases de hipótesis, la dimensión VC es aproximadamente lineal con el número de parámetros.
- Por tanto, el número de datos de entrenamiento necesario para entrenar crece linealmente con el número de parámetros de $\mathcal{H}$.

Dimensión VC para SVM

- ¿Qué pasa con lo SVM con kernels?
- ¿Mapeo a espacio de muchas dimensiones hace crecer la dimensión VC?
- Se ha demostrado que si $\|\underline{\mathbf{x}}^{(i)}\|_2 \leq R$, y solo aceptamos fronteras de separación con un margen geométrico mínimo γ, entonces

$$\text{VC}(\mathcal{H}) \leq \left\lceil \frac{R^2}{\gamma^2} \right\rceil + 1$$

- Por tanto, para SVM la dimensión VC sigue siendo baja

Resumen

Este documento ha sido elaborado con software libre incluyendo LaTeX, Beamer, GNUPlot, GNU/Octave, XFig, Inkscape, GNU-Make y Subversion en GNU/Linux