1. Máquinas de Aprendizaje

Héctor Allende

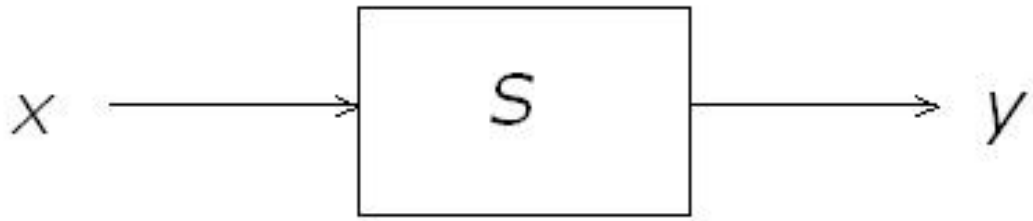
Universidad Técnica Federico Santa María

Febrero 2006

Agenda

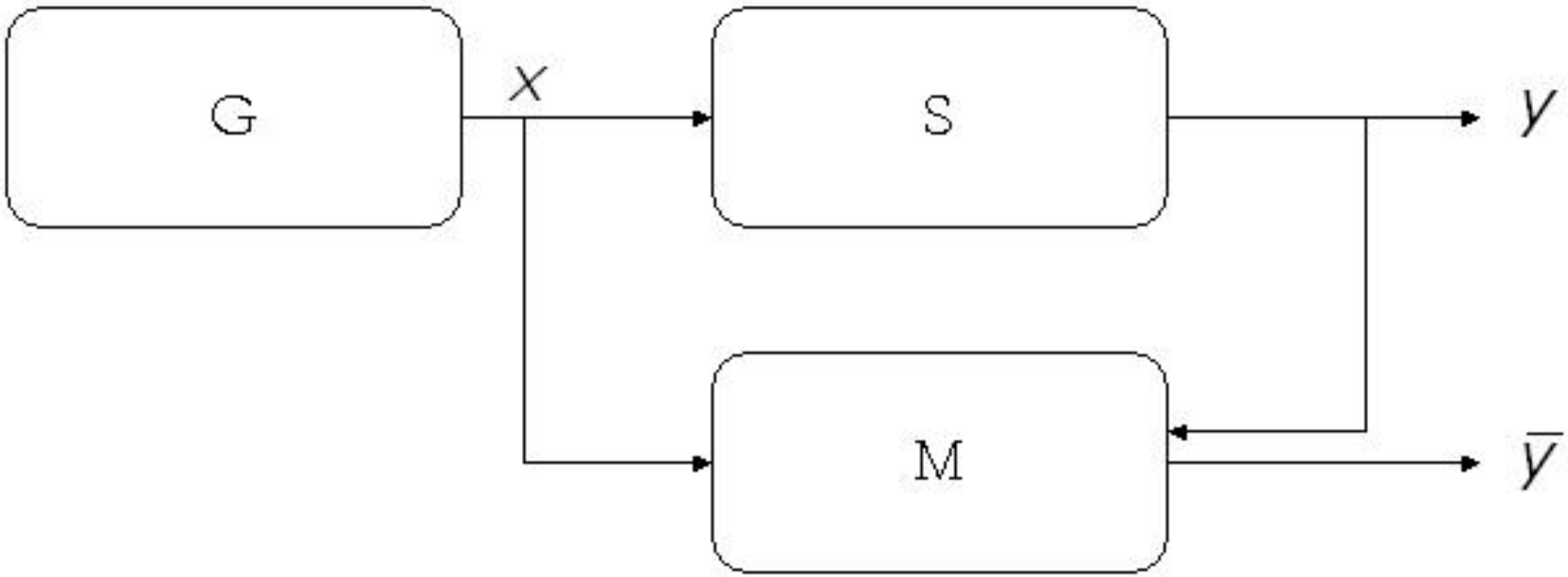
Problema General

- Materia Prima. Datos, observaciones, mediciones.
- Objetivo. Obtener descripciones de alto nivel de esos datos, relaciones, modelos, patrones.
- Problema de Inducción. ¿Cuál es el alcance de estas descripciones? ¿Son válidas más allá de los datos que permitieron construirlas?
- Supuesto: Existe una regularidad subyacente a las observaciones: Sistema Generador.



Problema General

- Identificar al Sistema: White-Box Models
- Imitar al Sistema: Black-Box Models



- ¿Identificar o imitar al sistema?

Modelo Estadístico del Problema

- Espacio de Observaciones $Z = X \times Y$ (... con una medida P que implementa el supuesto de "regularidad" subyacente a las observaciones)
- Espacio de Hipótesis H: Colección de todos los modelos seleccionables para explicar los datos
- Función de Pérdida $Q(f(x), y)$: ¿Cuál es el costo de responder $f(x)$ a una entrada x si el sistema responde con y?

Modelo Estadístico del Problema

- Aprendizaje: Elegir $F \in \Omega$ para minimizar el Riesgo asociado al modelo

$$E\left[Q(F(x), y)\right] = \int Q(F(x), y) dP(x, y) \tag{1}$$

No conocemos P!!, ¿Cómo elegir una hipótesis de H si no podemos computar (1)?

Principios de Inducción

- No conocemos P, pero sí un conjunto de ejemplos $D = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ que supondremos obtenidos i.i.d. de P.
- Funcional de Inducción: Criterio para elegir f sólo en base a la muestra D.

$$\begin{aligned} \hat{R} : H \times Z^n &\to \mathbb{R} \\ (f, D) &\mapsto R(f, D) \end{aligned} \tag{2}$$

- Elección Clásica: Funcional de Riesgo Empírico

$$\hat{R}^m(f, D) = \sum_{i=1}^{m} Q(f(x_i), y_i) \tag{3}$$

Principios de Inducción

- Se obtiene al reemplazar P por

$$\hat{P} = \sum_{i=1}^{m} \delta(x - x_i)\delta(y - y_i) \tag{4}$$

- Alternativas a este principio básico se pueden obtener al considerar otros estimadores de P, por ejemplo

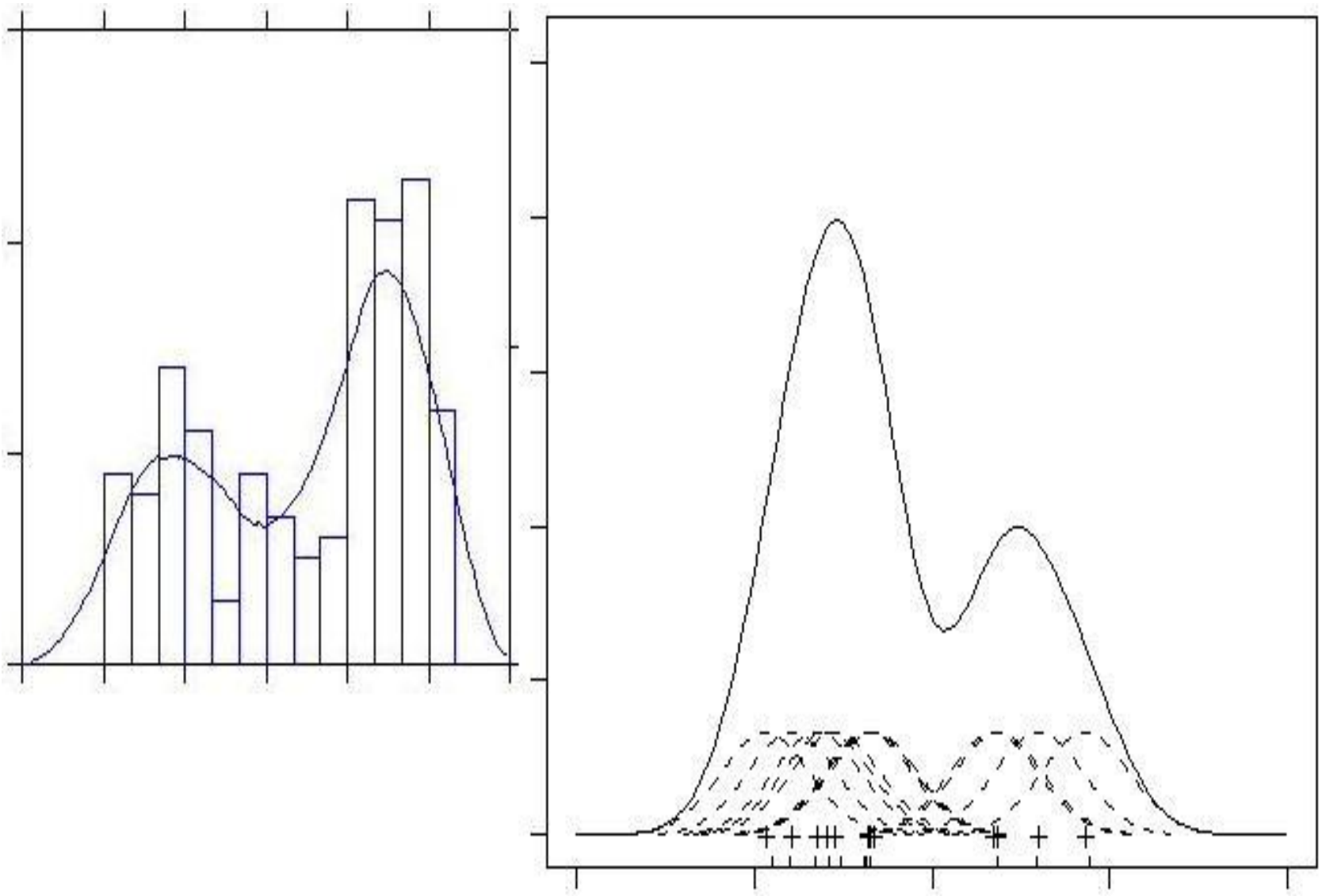
$$\hat{P} = \sum_{i=1}^{m} k(x - x_i, y - y_i) \tag{5}$$

con $k(\cdot, \cdot)$ un kernel centrado en $(0, 0)$.

- Obtenemos el denominado *Vicinal Risk*

$$\hat{R}(f, D) = \sum_{i=1}^{m} Q(f(x_i), y_i)k(x - x_i, y - y_i) \tag{6}$$

Principios de Inducción



Regresión

- ¿Qué principio de inducción es mejor?
- Antes de responder esta pregunta veamos como el modelo estadístico del aprendizaje puede servir para responder a problemas clásicos en estadística.
- Regresión: La relación entre X e Y es estocástica, i.e., dado un x existe un conjunto probable de respuestas y. Nos interesa estimar

$$r(x) = \int y dP(y|x) \tag{7}$$

denominada *función de regresión*

Regresión

- Consideremos el funcional de riesgo asociado a la función de pérdida cuadrática,

$$R(f) = \int (y - f(x))^2 dP(x, y) \tag{8}$$

- Entonces es posible mostrar que

$$R(f) = \int (y - r(x))^2 \, dP(x, y) + \int (f(x) - r(x))^2 \, dP(x)$$

- La primera parte cuantifica la varianza de y y la segunda la diferencia entre el modelo f y la curva de regresión $r(x)$.

- De esta forma, si

$$\int y^2 dP(x,y) < \infty \quad \int r^2(x) dP(x,y) < \infty$$

  el mínimo de $R(f)$ es la curva de regresión cuando $r(x) \in H$.
- Si $r(x) \notin H$, entonces f es función de H que resulta más cercana a $r(x)$ bajo la métrica $L_2(P)$.
- Además, si $\tilde{f}$ es tal que $R(\tilde{f}) - R(f) < \epsilon$, entonces la distancia entre f y $\tilde{f}$ es menor que $\sqrt{\epsilon}$.
- De esta forma, el problema de regresión (clásico) es un problema de aprendizaje con la función de pérdida cuadrática.

Clasificación

- Consideremos ahora el problema de distinguir entre un conjunto de categorías $w_1, w_2, \ldots, w_p$.
- Dado un conjunto de características x, supongamos que nuestro modelo f clasifica a x en la clase w_i.
- La probabilidad de error es entonces

$$e(f|x) = 1 - P(y = w_i|x) = \sum_{k \neq i} P(y = w_k|x) \tag{9}$$

- Definamos la función

$$L(y, f(x)) = \begin{cases} 0 & y = f(x) \\ 1 & y \neq f(x) \end{cases} \tag{10}$$

- Entonces la suma anterior es equivalente a

$$e(f|x) = \sum_k L(y, f(x))P(y|x) \tag{11}$$

- El error de clasificación esperado se obtiene integrando en el espacio de las características,

$$\begin{aligned} e(f) &= \int_X \sum_k L(y, \phi)P(y|x)P(x)dx \\ &= \int_X \sum_k L(y, \phi)P(y, x)dx \\ &= \int_{X\times Y} L(y, \phi)dP(x, y) \end{aligned}$$

- Esta expresión es exactamente el riesgo asociado a la función de pérdida $Q = L$ denominada *Misclassification Loss Function*.

Estimación de densidades

- Consideremos ahora el problema de estimar una función de densidad de probabilidad p_0 desde un conjunto H.
- La *entropía* para una función $p \in H$ se define como

$$R(p) = -\int p_0(x) \log p(x) dx = -\int \log p(x) dP(x) \tag{12}$$

Es decir, el riesgo asociado a la función de pérdida

$$Q(p) = -\log p(x) \tag{13}$$

- El mínimo del funcional anterior es el mismo que obtendríamos si sumamos a R una constante como $R(p_0)$,

$$R(p) = -\int p_0(x) \left(\log p(x) - \log p_0(x)\right) dx = -\int p_0(x) \log \frac{p(x)}{p_0(x)} dx$$

Estimación de densidades

- La minimización del funcional anterior es la base para la estimación de densidades en el approach estadístico clásico.
- Si aplicamos la desigualdad de Jensen y un poco de álgebra obtenemos la siguiente desigualdad para $R(p)$

$$R(p) \leq \log\left(1 - \left(\frac{1}{2}\int |p(x) - p_0(x)|dx\right)^2\right)$$

- Esta inecuación nos permite ver que el mínimo del funcional de riesgo se alcanza cuando $p(x)$ es la distribución verdadera $p_0(x)$.

Generalización

- Hemos visto que problemas clásicos pueden analizarse como problemas de minimización del funcional de riesgo.
- Como este funcional no puede ser computado en la práctica necesitamos un funcional empírico basado en los datos.
- Sin embargo necesitamos que el funcional permita elegir funciones que se comporten bien más allá de los ejemplos particulares que se utilizan para evaluarlo.

- *El funcional debe permitir generalizar*, i.e. el riesgo alcanzado con la hipótesis seleccionada debiera ser similar al riesgo mínimo que se puede obtener
- *El valor del funcional debe ser indicativo del riesgo real*, i.e., el valor del funcional empírico obtenido debiera ser también similar al mínimo que se puede obtener.

Generalización y Consistencia

- Para formalizar esta idea, sea f_m la hipótesis que minimiza $\hat{R}$ en una muestra de tamaño m y f_0 la hipótesis que minimiza el riesgo real R.

Consistencia (Débil)

Diremos que el principio de inducción $\tilde{R} : \Omega \longrightarrow \mathbb{R}$ es *consistente* si y sólo si se satisfacen las siguientes condiciones

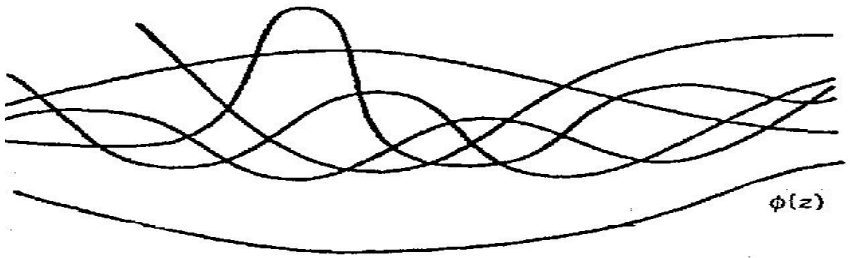
$$R(f_m) \xrightarrow[m]{P} \quad R(f_0) \tag{14}$$

$$\tilde{R}(f_m) \xrightarrow[m]{P} \quad R(f_0) \tag{15}$$

- Hablaremos de "consistencia universal" si $\hat{R}$ es consistente independiente de la distribución $P(x, y)$ de los ejemplos.

Generalización y Consistencia

- Supongamos que en el espacio Ω existe una función ϕ tal que $\forall(x, y)$, $Q(y, \phi(x)) < Q(y, f(x))$, $\forall f \in \Omega$.
- Es evidente que cualquier principio de inducción $\tilde{R}$ basado en cualquier muestra D encontrará $\phi(x)$ como solución al problema.
- Esto nos indica que la noción de consistencia es demasiado débil y necesita un poco de precisión.



Consistencia Fuerte

Consistencia (Estricta)

Diremos que un principio de inducción $\tilde{R}$ es estrictamente consistente en H si en cualquier $\lambda(c) \neq \varnothing$

$$\lambda(c) := \{g \in H : R(g) \geq c\} \tag{16}$$

se satisface

$$\tilde{R}(f_m^c) \xrightarrow[m]{P} R(f_0^c) \tag{17}$$

donde f_m^c es la función que minimiza $\tilde{R}$ en el subconjunto $\lambda(c)$ y f_0^c es la función que minimiza R en el mismo subconjunto.

- Notemos que claramente la consistencia estricta implica la consistencia débil

Teorema Clave de la Teoría VC

- Uno de los resultados claves de la *Teoría Estadística del Aprendizaje* es el que relaciona la consistencia del principio del riesgo empírico con el estudio de un proceso empírico del "peor caso".

Proceso Empírico de una Cola

Se define como la secuencia $(\xi_+^m)_m$, $m \in \mathbb{N}$ donde

$$\xi_+^m = \sup_{f \in H} \left(R(f) - \hat{R}^m(f) \right)$$

y $\hat{R}^m = 1/m \sum_{i=1}^m Q(f(x_i), y_i)$

- Hablamos del "peor caso" pues para cada tamaño muestral m tomamos el supremo de las diferencias entre R y $\hat{R}^m$.

Teorema Clave de la Teoría VC

Y ahora el resultado clave ...

Teorema de Equivalencia (Vapnik, Chervonenkis)

Sea $R(f)$ absolutamente acotado en H. Entonces las siguientes condiciones son equivalentes:

1. $\hat{R}^m$ es estrictamente consistente en H.
2. El proceso empírico de una cola converge a cero en probabilidad, es decir

$$P\left(\xi_+^m > \epsilon\right) \underset{m\to\infty}{\longrightarrow} 0 \tag{18}$$

- Si una de las condiciones es válida independiente de la distribución $P(x, y)$ de los ejemplos, la otra también es universal.

Teorema Clave de la Teoría VC

- ¿Bajo que condiciones el proceso de una cola converge a cero?. Consideremos un caso simple en que Q toma valores en $\{0, 1\}$ (clasificación) y H es finito.
- La desigualdad de Chernoff puede aplicarse sobre cada función de H para concluir que

$$\begin{aligned} P\left(\xi_+^m > \epsilon\right) &\leq 2Ne^{-2\epsilon^2 m} \\ &\leq 2e^{\left(\frac{\ln N}{l} - 2\epsilon^2\right)m} \end{aligned}$$

- Esto sugiere que si

$$\frac{\ln N}{m} \underset{m\to\infty}{\longrightarrow} 0$$

el proceso de una cola convergerá a cero en probabilidad.

Capacidad y Dilema Bias-Varianza

- El resultado anterior sugiere que el tamaño del espacio de soluciones H debe mantenerse bajo control para obtener consistencia y una buena generalización.
- Conceptualmente esta idea es la misma que expresa el clásico dilema Sesgo-Varianza en estadística.
- El Sesgo de un estimador f_m que se obtiene de una muestra de ejemplos D se define como

$$Bias^2(f_m) = E_P\left(E_T\left[f_m(x)\right] - f_0(x)\right)^2$$

donde T es la distribución conjunta de los ejemplos.

- La varianza del estimador se define como

$$Var(f_m) = E_P E_T\left(f_m(x) - E_T\left[f_m(x)\right]\right)^2$$

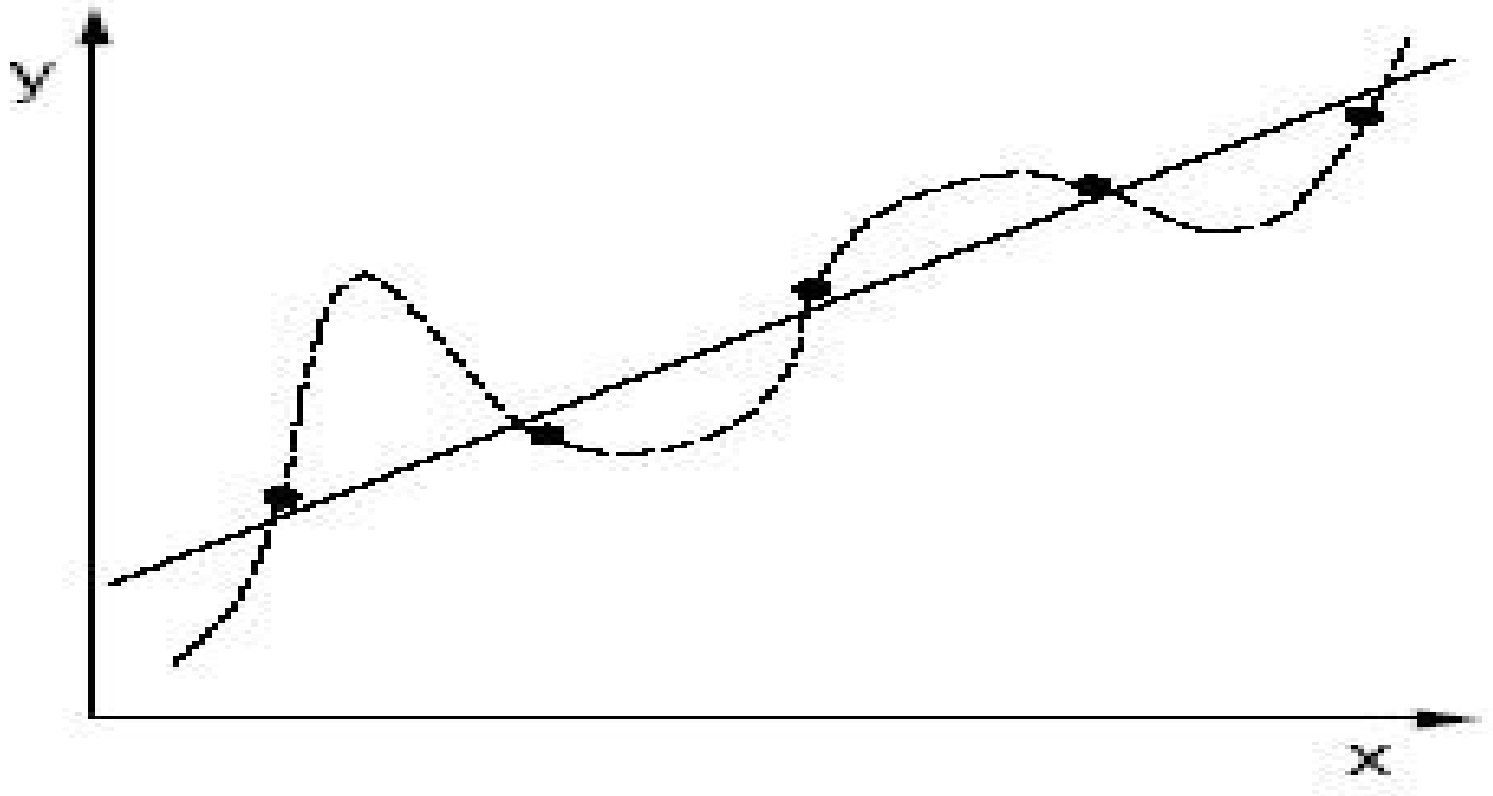
Capacidad y Dilema Bias-Varianza

- A medida que aumentamos el tamaño del espacio de soluciones es más probable encontrar una función que modele perfectamente el conjunto de ejemplos.
- Por otro lado, si tenemos muchas soluciones, pequeños cambios en el conjunto de ejemplos darán origen a una solución diferente. Es decir, la varianza esperada tiende a aumentar si el espacio se hace más grande.
- Intuitivamente estimadores inestables tienden a no generalizar adecuadamente i.e. sobreajustan.

Al modificar la complejidad estructural del espacio de soluciones existe un compromiso entre sesgo y varianza ó bien entre ajuste y generalización.

Capacidad y Dilema Bias-Varianza

Consideremos el siguiente ejemplo:



- ¿Qué aproximación es preferible?

Capacidad y Dilema Bias-Varianza

- Un ejemplo de control de capacidad se da en la interpolación con splines cúbicos. En este caso se favorecen los modelos "simples" utilizando un término inversamente proporcional a la segunda derivada del estimador,

$$\hat{R}(f) = \sum_{i=1}^{m} (y_i - f(x_i))^2 + \lambda \int \left[f''(x)\right]^2 dx$$

- Otro ejemplo clásico son las medidas para selección de modelos denominadas *BIC* ó *AIK* (Criterio de Akaike) que son inversamente proporcionales a la varianza muestral.
- El tema es cómo medir adecuadamente la "capacidad" del espacio de soluciones. Claramente la cardinalidad no es suficiente pues en general trabajaremos con espacios infinitos.

Medidas de Capacidad

- El funcional de riesgo empírico genérico no toma en cuenta ninguna medida estructural. Por lo tanto su capacidad para generalizar debe depender de características del espacio de hipótesis H.
- El tema es cómo medir adecuadamente la "capacidad" del espacio de soluciones. Claramente la cardinalidad no es suficiente pues en general trabajaremos con espacios infinito-dimensionales.
- Observación Clave: No interesa directamente el tamaño del espacio sino las diferentes configuraciones que se alcanzan.
- Dos funciones que siempre reportan el mismo riesgo son idénticas aunque paramétricamente no lo sean.

Medidas de Capacidad

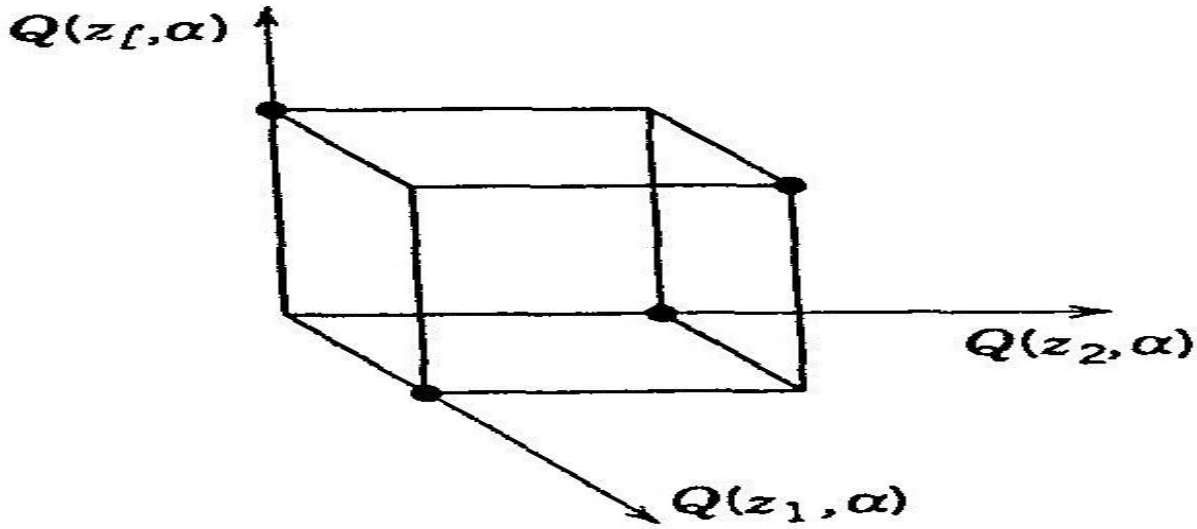
- La observación anterior, motiva la definición de una medida de capacidad que tome en cuenta los valores que puede tomar la función $Q(f(x), y)$ en el espacio H.
- Dado un conjunto de ejemplos $D_m = \{(x_i, y_i), i = 1, \ldots, m\}$ y una función $f \in H$ definamos,

$$q(f) = (Q(f(x_1), y_1), \ldots, Q(f(x_m), y_m))$$

- De esta forma, $q(H) = \{q(f), f \in H\}$.
- En el caso más simple en que Q toma valores en $\{0, 1\}$, $\#q(H)$ es finita ($\leq 2^m$) y parece una medida apropiada de capacidad.

Medidas de Capacidad

- Gráficamente $\#q(H)$ es el número de vértices diferentes que se inducen sobre el hipercubo $q(H)$ que se forma tomando en cada eje un ejemplo $z_i = (x_i, y_i)$.



Entropía

- Sobre $\#q(H)$ podemos introducir las siguientes definiciones de capacidad.

Entropía de un Espacio de Hipótesis

Sea $P(x, y)$ la distribución del conjunto de ejemplos D. La *entropía aleatoria* de un espacio de hipótesis H se define como

$$\mathcal{E}^H(D) = \ln\left(\#q(H)\right) \tag{19}$$

Se define además la *Entropía* de H

$$\mathcal{E}^H(m) = E_P\left(\mathcal{E}^H(D)\right) \tag{20}$$

Capacidad versus Generalización

Teorema

Para que el proceso de una cola asociado a un funcional de inducción $\hat{R}$ converja en probabilidad a cero es suficiente que se verifique

$$\frac{\mathcal{E}^H(m)}{m} \underset{m\to\infty}{\longrightarrow} 0 \tag{21}$$

- Este teorema relaciona la habilidad de generalización de una máquina de aprendizaje -que implemente el riesgo empírico- con una medida específica de capacidad.

Generalización de las Medidas

- La búsqueda de condiciones que sean además de suficientes sean necesarias y que sean válidas para funciones Q con valores en $\mathbb{R}$ nos llevan a explorar otros conceptos de capacidad.

ϵ-net y Número de Cubrimiento

Sea A un conjunto dotado de una métrica ρ. Una ϵ-net de A es un conjunto Λ tal que $\forall f \in A$, existe $\tilde{f} \in \Lambda$ que satisface

$$\rho(f, \tilde{f}) < \epsilon$$

Denominamos *ϵ-net minimal* a aquella ϵ-net $\tilde{\Lambda}$ tal que $\tilde{\Lambda} \subset \Lambda$ para toda ϵ-net Λ. Llamaremos número de cubrimiento (covering number) $\mathcal{N}(\epsilon, A, \rho)$ a la cardinalidad de este conjunto minimal.

ϵ-Entropía

- Consideremos el conjunto $q(H)$ formado por los vectores $q(f) = (Q(f(x_1), y_1), \ldots, Q(f(x_m), y_m)), \forall f \in H$.

ϵ-entropía

Sea $P(x, y)$ la distribución del conjunto de ejemplos D. La *ϵ-entropía aleatoria* de un espacio de hipótesis H se define como

$$\mathcal{E}^H(\epsilon, D) = \ln\left(\mathcal{N}(\epsilon, q(H), l^\infty)\right) \tag{22}$$

donde l^∞ es la métrica del supremo. Se define además la *ϵ-entropía* de H

$$\mathcal{E}^H(\epsilon, m) = E_P\left(\mathcal{E}^H(\epsilon, D)\right) \tag{23}$$

- El siguiente teorema cierra el problema de consistencia del riesgo empírico utilizando las medidas introducidas.

Teorema de Consistencia del Riesgo Empírico

Para que obtener convergencia en probabilidad del proceso de una cola asociado a un funcional de inducción $\hat{R}$ es suficiente y necesario que para todo ϵ, δ, σ exista un σ-cubrimiento Θ de H tal que

$$\lim_{m \to \infty} \frac{\mathcal{E}^{\Theta}(\epsilon, m)}{m} < \delta \tag{24}$$

Θ es un σ-cubrimiento de H si $\forall f \in H \exists \tilde{f} \in Co$ tal que

$$Q(f(x), y) \geq Q(\tilde{f}(x), y)$$
$$R(f) - R(\tilde{f}) < \sigma$$

Consistencia Universalmente Válida

- La noción de entropía depende de la distribución de probabilidad $P(x, y)$ asociada a los ejemplos en D.
- Para asegurar la consistencia del riesgo empírico de manera independiente de la medida de probabilidad es necesario introducir una cota para la entropía, que denominaremos *función de crecimiento* (*growth function*) del espacio de hipótesis H,

$$\mathcal{G}(\epsilon, H, m) = \ln \sup_{D} \mathcal{N}(\epsilon, q(H), l^{\infty})$$

- Para hacer universales los teoremas de consistencia es suficiente reemplazar la entropía $\mathcal{E}^{H}(\epsilon, m)$ por $\mathcal{G}(\epsilon, H, m)$.

Dimensión VC

Teorema

Supongamos que $q(H)$ esta formada por funciones que toman valores en $\{0,1\}$. Entonces

$$\mathcal{G}(H,m) = m \ln m$$

ó bien existe un $h \in \mathbb{N}$ tal que

$$\mathcal{G}(H,m) \begin{cases} = m \ln m & m \leq h \\ \leq h\left(1 + \ln \frac{m}{h}\right) & l > h \end{cases}$$

- Si ocurre el primer caso diremos que h es infinita.
- El número h se conoce como Dimensión VC

Dimensión VC

- Recordemos que el principio del riesgo empírico es universalmente consistente en el conjunto de indicatrices H si se verifica

$$\lim_{m\to\infty} \frac{\mathcal{G}(H,m)}{m} = 0$$

- Notemos que si H tiene dimensión VC finita h, entonces tenemos,

$$\lim_{m\to\infty} \frac{\mathcal{G}(H,m)}{m} < \lim_{m\to\infty} \frac{h\left(1+\ln\frac{m}{h}\right)}{m} = 0$$

Si el espacio de hipótesis tiene dimensión VC finita se verifica universalmente la consistencia del riesgo empírico. Sin embargo no se trata de una condición necesaria.

Cotas VC Universales

Teorema: Cota Aditiva para el Riesgo Empírico

Supongamos que la dimensión VC h asociada a un espacio de hipótesis H es finita. Entonces,independiente de la distribución de los ejemplos,

$$P\left(\sup_{f\in H}\left|R(f)-\hat{R}^m(f)\right|>\epsilon\right)<4\exp\left(\frac{h\left(1+\ln(2m/h)\right)}{m}-(\epsilon-1/m)^2\right)m$$

Equivalente podemos decir, que con probabilidad $1-\eta$

$$R(f_m)<\hat{R}^m(f_m)+\sqrt{\mathcal{L}(m)}+\frac{1}{m}$$

donde f_m es el mínimo del riesgo empírico $\hat{R}^m$

$$\mathcal{L}(m)=4\frac{h\left(\ln(2m/h)+1\right)-\ln\eta/4}{m}$$

Teorema: Cota Relativa para el Riesgo Empírico

Supongamos que la dimensión VC h asociada a un espacio de hipótesis H es finita. Entonces,

$$P\left(\sup_{f\in H}\frac{R(f)-\hat{R}^m(f)}{\sqrt{R(f)}}>\epsilon\right)<4\exp\left(\frac{h\left(1+\ln(2m/h)\right)}{m}-\frac{\epsilon^2}{4}\right)m$$

Equivalente podemos decir, que con probabilidad $1-\eta$

$$R(f_m)<\hat{R}^m(f_m)+\frac{\mathcal{L}(m)}{2}\left(1+\sqrt{1+\frac{4\hat{R}^m(f_m)}{\mathcal{L}(m)}}\right)$$

donde f_m es el mínimo del riesgo empírico $\hat{R}^m$.

Dimensión VC para funciones reales

- La dimensión VC para el caso de un espacio $q(H)$ toma valores reales se puede definir introduciendo un conjunto auxiliar $\Theta(H)$ de funciones que tomen valores en $\{0, 1\}$.

$$\Theta(H) = \{\theta(Q(f(x), y) - \beta), f \in H, \beta \in \mathbb{R}\}$$

  donde θ es una indicatriz centrada en 0.

- La función de crecimiento de H se define en este caso como en el caso de funciones valuadas en $\{0, 1\}$ pero sobre el conjunto $\Theta(H)$. Lo mismo con la dimensión VC.
- Este truco permite extender todos los teoremas al caso general de funciones Q reales.

Observaciones acerca de la teoría VC

- Si bien el truco anterior permite obtener resultados elegantes para explicar las condiciones en que el riesgo empírico es consistente, éste no incorpora ninguna información acerca de las escala de las observaciones.
- En el caso de clasificación dicha información de escala no es importante pues los valores que toman las hipótesis $f(x)$ son sólo una referencia para distinguir entre clases.
- En el caso de regresión sin embargo la pérdida de la información de escala hace que los resultados de la teoría VC no sean útiles en la práctica. Por ejemplo, las cotas para el riesgo empírico suelen ser muy poco *tight* como para ser informativas.

Regularización y Control de Capacidad

- Los resultados de la teoría VC afirman la idea de que una máquina de aprendizaje que generaliza, debe mantener controlada la capacidad del espacio de soluciones H.

Las cotas para el riesgo obtenidas de la teoría VC tienen la forma genérica

$$R(f) \leq \tilde{R}^m(f) + \Lambda(H)$$

donde $\hat{R}^m$ es el riesgo empírico y $\Lambda(H)$ es una medida de la complejidad estructural del espacio de soluciones.

- $\Lambda(H)$ es una medida global en el espacio de soluciones. No depende de una hipótesis candidata en particular. Más aún es independiente de los datos.
- El desarrollo de un principio adaptivo de aprendizaje a partir de estos resultados es difícil.

Regularización y Control de Capacidad

- La idea clave en regularización es la restricción del espacio de soluciones mediante la incorporación de un funcional $\Omega(f)$ que penalize hipótesis complejas.
- El funcional de riesgo empírico regularizado tiene la forma

$$\tilde{R}^{reg}(f) = \hat{R}^m(f) + \lambda\Omega(f) \tag{25}$$

  donde $\Omega(\cdot)$ se denomina *regularizador* y el correspondiente λ *parámetro de regularización*.
- El parámetro λ controla el tradeoff entre ajuste a los datos ($\hat{R}^m(f)$) y complejidad estructural ($\Omega(f)$).
- Hipótesis complejas son admisibles siempre que los datos lo requieran.

Regularización y Control de Capacidad

- Un ejemplo clásico (en estadística bayesiana) consiste en considerar un regularizador proporcional a la norma de la primera derivada

$$\hat{R}^{reg}(f) = \hat{R}^{m}(f) + \lambda \|\delta_x f\|^2$$

- De entre dos funciones con el mismo riesgo empírico, se prefiere la que sea más suave.
- Otro ejemplo se da en redes neuronales (*weight decay neural networks*)

$$\hat{R}^{reg}(f) = \hat{R}^{m}(f) + \lambda \|w\|^2$$

- En este caso w es el vector de pesos de la red. $\|w\|^2$ intenta restringir la norma de la hipótesis $\|f\|^2$.

Problemas Inversos

- Consideremos dos espacios de funciones M y N y un operador $A : M \longrightarrow N$ que asigna a cada $f \in M$ un único $F \in N$.
- Supongamos que queremos estimar f, pero sólo podemos observar F. Si conocemos A, el problema se reduce a resolver la ecuación,

$$Af = F \tag{26}$$

- Denominaremos este tipo de problemas Problemas Inversos.
- Por ejemplo, la estimación de densidades $f(t)$ a partir de un conjunto de ejemplos es un problema inverso

$$\int_{-\infty}^{x} f(t)dt = F(x)$$

donde $F(x)$ es la distribución de los ejemplos.

Problemas Inversos

- Para que el proceso se estimación tenga sentido, nos interesa que dos soluciones próximas en el espacio observable N tengan preimágenes cercanas en el espacio de las soluciones M.
- Esta idea se captura técnicamente en la noción de continuidad la cual requiere que M y N sean espacios métricos. El operador A es continuo

Continuidad de Operadores

Sean (X, ρ_1) y (Y, ρ_2) dos espacios métricos y $C : X \rightarrow Y$ un operador. C se dice continuo si y sólo si $\forall \epsilon, \forall x \in A \exists \sigma > 0$ tal que

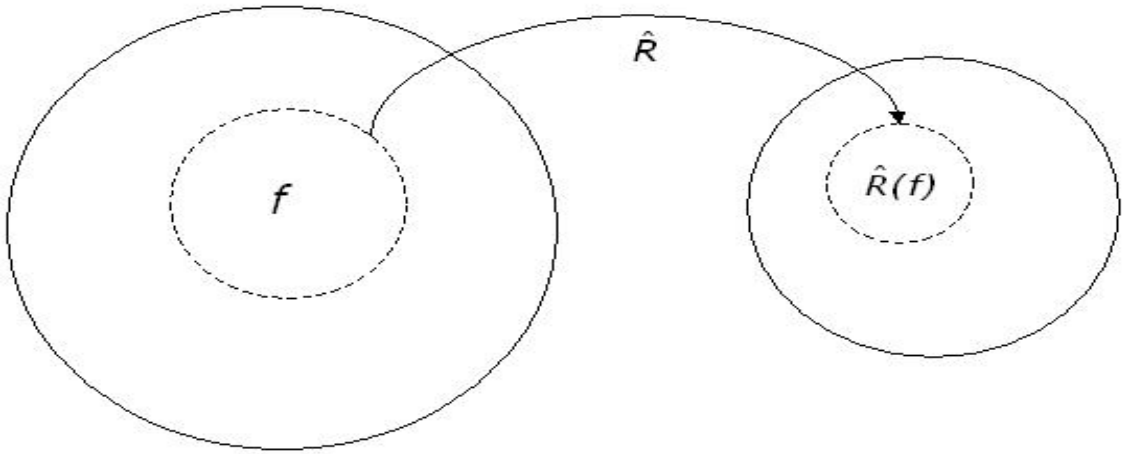
$$CB_{\rho_1}(x, \sigma) \subset B_{\rho_2}(Ax, \epsilon)$$

donde $B_{\rho_1}(x, \sigma)$ es la bola (abierta) de radio σ en la métrica ρ_1 y análogamente $B_{\rho_2}(Ax, \epsilon)$.

- Claramente nos interesa la continuidad del operador inverso A^{-1}.

Aprendizaje como Problema Inverso

- El problema de estimar una función desde un conjunto de ejemplos puede verse como un problema inverso.
- El operador A es en este caso un funcional de inducción $\hat{R}$ que mapea cada función de un espacio de hipótesis H a un valor único $\hat{R}(f)$ en $\mathbb{R}$.



Nos interesa que si dos hipótesis tienen valores similares en el funcional de inducción, éstas estén próximas en el espacio de hipótesis pues esto garantiza que las funciones son similares más allá de los ejemplos particulares.

Aprendizaje como Problema Inverso

- Una visión alternativa se obtiene al considerar A como el operador que toma una función f del espacio de hipótesis y genera una muestra de ejemplos $D(f) = \{(x_i, f(x_i), i = 1, \ldots, m)\}$.
- En este caso $X_m = \{x_i; i = 1, \ldots, m\}$ es un parámetro del operador A, encargado de evaluar f en X_m.
- El problema es estimar f desde una muestra particular.

Nos interesa que dos conjuntos de ejemplos similares generen hipótesis similares. El operador A^{-1} que mapea un conjunto de ejemplos a una función en el espacio de soluciones es nuestro algoritmo. Nos interesa que el algoritmo sea estable.

Buen y Mal Condicionamiento

Buen Condicionamiento en el sentido de Hadamard

Un problema inverso $Af = F$ con $f \in M$ y $F \in N$ se dice *bien condicionado en el sentido de Hadamard* (*well-posed*) ssi

1. Para todo F, f existe y es única
2. El problema es estable, i.e., A^{-1} es continuo

Buen Condicionamiento en el sentido de Tikhonov

Un problema inverso $Af = F$ con $f \in M$ y $F \in N$ se dice *bien condicionado en el sentido de Tikhonov* ssi $\exists M' \subset M$ tal que

1. Para todo $F \in N' = AM'$, f existe, es única y esta en el *espacio de correctitud* M'
2. El problema es estable, i.e., A^{-1} es continuo en M' y N'.

Buen y Mal Condicionamiento

- Notemos que si el problema está bien condicionado, entonces para cualquier secuencia $F_1, F_2, \ldots, F_m$ convergente a F la secuencia de las preimágenes $f_1, f_2, \ldots, f_m$, $f_i = A^{-1}F$ converge a $f = A^{-1}F$.
- Condiciones suficientes bajo las cuales tenemos un problema bien condicionado son las siguientes:

Teorema Clave de Regularización

Sea $M' \subset M$ un espacio métrico compacto y $A : M \to N$ un operador continuo. Entonces $A^{-1} : N' \to M'$ es continuo en $N' = AM'$.

Método de Regularización

- Introducido por Tikhonov hacia 1963 para resolver el problema inverso $Af_0 = F_0$, $f_0 \in M, F_0 \in N$.
- La idea consiste en restringir las soluciones candidatas f a un subespacio compacto de M.
- Para ello se considera un funcional semicontinuo $W : M \to \mathbb{R}$ denominado regularizador.

Regularizador

$W : M \to \mathbb{R}$ se denomina regularizador si es semicontinuo y satisface las propiedades:

1. W está definido para f_0.
2. En su dominio de definición, $W(f) \geq 0$
3. Los conjuntos
$$M_c := \{f \in M : W(f) \leq c\}$$
son todos compactos.

Método de Regularización

- Para resolver el problema inverso, lo tradicional es minimizar en el espacio M, la norma de la diferencia

$$\tilde{R}(f, F_0) = \rho_2(Af - F_0)$$

donde ρ_2 es la norma del espacio observable N.

- El problema es que si en vez de observar F_0 observamos F_δ muy cercano a F_0, digamos

$$\rho_2(F_\delta - F) \leq \delta$$

la solución f_δ que minimiza $\tilde{R}(f, F_\delta)$ no resulta cercana a f_0, a menos que el problema sea bien condicionado.

Método de Regularización

- Tikhonov introduce en cambio el funcional

$$\hat{R}_{\lambda}^{reg}(f, F_0) = \rho_2^2(Af - F_0) + \gamma W(f)$$

con $\gamma > 0$.

- Si minimizamos ahora, $\hat{R}_{\lambda}^{reg}(f, F_\delta)$ ¿La solución que se obtiene f_δ^γ, es cercana a f_0?
- Más aún, supongamos que tenemos secuencias F_δ convergentes a F_0 ¿Las correspondientes secuencias f_δ^γ son convergentes a f_0?
- La respuesta dependerá de que tan cercano sea F_δ a F_0 y de cómo se elija el parámetro γ.

Método de Regularización

Teorema

Sean (M, ρ_1),(N, ρ_2) espacios métricos. Supongamos que en vez de F_0 observamos aproximaciones $F_\delta \in N$ tal que $\rho_2(F_0 - F_\delta) \leq \delta$. Supongamos que el parámetro de regularización es tal que

$$\gamma(\delta) \underset{\delta \to 0}{\longrightarrow} 0$$

y además,

$$\lim_{\delta \to 0} \frac{\delta^2}{\gamma(\delta)} \leq r < \infty$$

Entonces, los elementos $f_\delta^{\gamma(\delta)}$ que minimizan los funcionales $\tilde{R}_{\gamma(\delta)}(f, F_\delta)$ convergen a la solución exacta $f_0 = A^{-1}F_0$ a medida que $\delta \to 0$.

Método de Regularización

- La demostración del teorema anterior se basa (en parte) en la siguiente desigualdad

$$W(f_\delta^{\gamma(\delta)}) \geq W(f) + \frac{\delta^2}{\gamma(\delta)}$$

- Como los conjuntos $W(f) \leq c$ son compactos $\forall c$, la minimización de R_γ^{reg} considerando un F_δ cercano a F_0 genera $f_\delta^{\gamma(\delta)}$ cercanos a f_0.
- Las condiciones impuestas sobre $\gamma(\delta)$ permiten concluir el resultado más fuerte de que secuencias de F_δ's convergentes en N generan secuencias $f_\delta^{\gamma(\delta)}$'s convergentes en M.

Método de Regularización

- Supongamos que M es un espacio de Hilbert (e.g., en métodos de kernel). Una elección típica es en este caso

$$W^H(f) = \|f\|^2 = \langle f, f \rangle$$

- ¿Es W^H un regularizador válido?. La respuesta es negativa dado que los conjuntos $M_c = \{f, W^H(f) \leq c\}$ son sólo débilmente compactos.
- Sin embargo las propiedades de un espacio de Hilbert permiten establecer

Teorema

Sea M un espacio de Hilbert y $W(f) = \langle f, f \rangle$. Entonces si γ_0 satisface las condiciones del teorema anterior con $r = 0$, los elementos $f_\delta^{\gamma(\delta)}$ convergen a f_0.

- Desde el punto de vista bayesiano (McKay, 1996) el problema de aprendizaje se entiende como la estimación de la distribución *a-posteriori* de las hipótesis $p(h|(X,Y))$ dado un conjunto de observaciones u ejemplos $(X,Y) = \{(x_1, y_1), i = 1, \ldots, m\}$.
- Usando la regla de Bayes podemos encontrar que

$$p(h|X,Y) = \frac{p(X,Y|h)p(h)}{p(Y|X)}$$

- La característica central de la estimación bayesiana es que asume cierto conocimiento previo de los datos y del tipo de relaciones funcionales que podemos encontrar.
- Este conocimiento previo se recoge imponiendo una estructura particular a $p(h)$ y a $p(X,Y|h)$.

- Desde el punto de vista bayesiano (McKay, 1996) el problema de aprendizaje se entiende como la estimación de la distribución *a-posteriori* de las hipótesis $p(h|(X,Y))$ dado un conjunto de observaciones u ejemplos $(X,Y) = \{(x_1, y_1), i = 1, \ldots, m\}$.
- Usando la regla de Bayes podemos encontrar que

$$p(h|X,Y) = \frac{p(X,Y|h)p(h)}{p(Y|X)}$$

- La característica central de la estimación bayesiana es que asume cierto conocimiento previo de los datos y del tipo de relaciones funcionales que podemos encontrar.

Este conocimiento previo se recoge imponiendo una estructura particular a $p(h)$ denominado *a-priori* y a $p(X,Y|h)$ denominada *verosimilitud*.

- El *a-posteriori* puede verse como una corrección del *a-priori* distribucional tomando en cuenta la información que nos entregan las observaciones.
- Supongamos que los datos se generan según

$$y_i = f(x_i) + \xi_i \Leftrightarrow \qquad y - f(x_i) = \xi_i$$

- En este caso, es razonable suponer que

$$p(x_i, y_i) = p(y_i - f(x_i)) = p(\xi_i)$$

- Si el conjunto de datos se genera de manera independiente e idénticamente distribuida, obtenemos que

$$p(X, Y|h) = \prod_{i=1}^{m} p(\xi_i)$$

- Dado el *a-posteriori* $p(h|X,Y)$ y una determinada entrada x la predicción de y corresponde a estimar

$$p(y|X,Y,x) = \int_H p(y|h,x)p(h|X,Y)dh$$

- Aproximación MAP (*Maximum a Posteriori*). Muchas veces aproximación de este tipo de integrales es intratable. Una aproximación razonable consiste en elegir la hipótesis de H que maximiza $p(h|X,Y)$ y utilizar esta hipótesis para generar la salida $y = f(x)$.
- Maximizar $p(h|X,Y)$ es equivalente a minimizar $-\ln p(h|X,Y)$, que a su vez resulta equivalente a minimizar

$$\tilde{R}^{bay}(h) = -\ln p(X,Y|h) - \ln p(h) = \sum_{i=1}^{m} -\ln p(y - h(x_i)) - \ln p(h)$$

- Notemos la similitud de la función objetivo anterior y el riesgo empírico regularizado

$$\tilde{R}^{bay}(h) \propto R_{emp}(h) + \lambda\Omega(h)$$

- Consideremos el siguiente *a-priori* gaussiano sobre H con parámetro de escala σ_h

$$p(h) = \frac{1}{K_h} e^{-\frac{\|h\|^2}{\sigma_h}}$$

- Supongamos además que el ruido asociado al modelo es gaussiano con parámetro de escala σ_ξ,

$$p(\xi) = \frac{1}{K_\xi} e^{-\frac{\xi^2}{\sigma_\xi}}$$

- Entonces $\tilde{R}^{bay}(h)$ tiene la forma,

$$\begin{aligned}
\tilde{R}^{bay}(h) &= \sum_{i=1}^{m} -\ln p(y - h(x_i)) - \ln p(h) \\
&= \sum_{i=1}^{m} \frac{1}{\sigma_\xi}(y_i - f(x_i))^2 + \frac{1}{\sigma_h}\|h\|^2 \\
&\propto \frac{1}{m}\sum_{i=1}^{m}(y_i - f(x_i))^2 + \frac{\sigma_\xi}{m\sigma_h}\|h\|^2
\end{aligned}$$

- Notemos que $\tilde{R}^{bay}(h)$ coincide con el riesgo regularizado R_γ^{reg} si $\gamma = \frac{\sigma_\xi}{m\sigma_h}$.

- El *a-priori* $p(h)$ sobre el espacio de las soluciones tiene el rol de regularizar la información acerca del error que aportan las observaciones.
- Seleccionar un regularizador es equivalente a seleccionar un *a-priori* sobre H.
- Esta "visión bayesiana" también nos previene de que el término de error depende del modelo de ruido que estemos asumiendo, quizá implícitamente.

Temas Pendientes

- Teoría Algorítmica del Aprendizaje
- Generalización, Estabilidad y Robustez
- Funciones de Pérdida y Modelos de Ruido