TRATAMIENTO DIGITAL DE SEÑALES
Ingeniería de Telecomunicación (4º, 2º c)

Unidad 11ª: Sobre los costes

Aníbal R. Figueiras Vidal
Jesús Cid Sueiro
Ángel Navia Vázquez

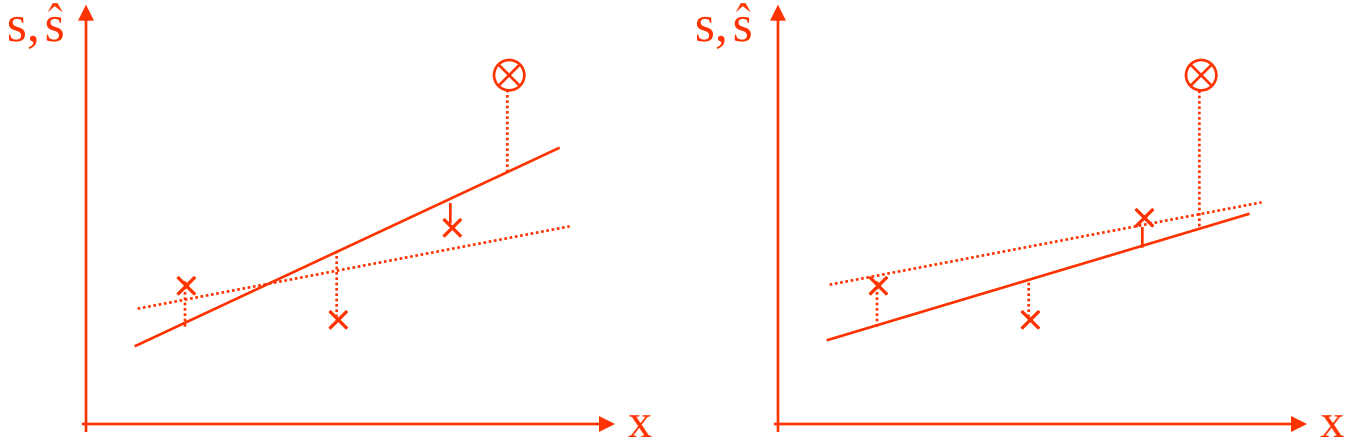
Área de Teoría de la Señal y Comunicaciones
Universidad Carlos III de Madrid

A. Costes para estimación

* En principio, puede emplearse como coste cualquier medida de similitud
  - en diseño analítico, entre s y $\hat{s}$
  - en el entrenamiento máquina, entre d y o.
* Su elección puede venir determinada por el propósito del diseñador (según el problema): pero en muchas situaciones no hay motivos claros para preferir un coste a otro.
* No obstante, hay aspectos del problema a resolver que pueden orientar sobre la elección.

Discusión

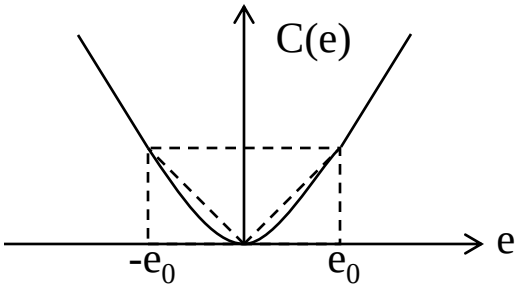
D: Como se sabe, una muestra fuera de margen ("outlier") es la que se aleja significativamente del comportamiento implicado por todas las demás. Valore el efecto de la aparición de una en un problema de regresión lineal según se pretenda minimizar un coste cuadrático o un coste absoluto.



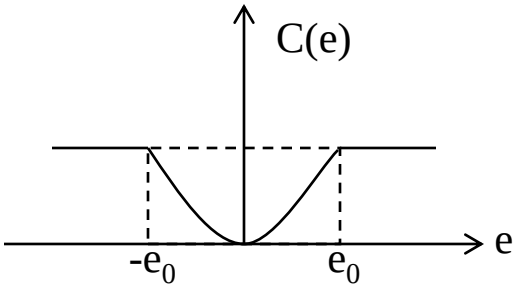
Supuesta una recta de regresión "teórica" (a trazos), la muestra fuera de margen desplaza su versión estimada (línea continua) más sensiblemente cuando se emplea un coste cuadrático, ya que el valor del error correspondiente al cuadrado será muy grande.

Esta sensibilidad del coste cuadrático a las muestras fuera de margen supone una limitación para su uso. Para remediarlo, puede procederse a censurar muestras que tomen valores atípicos (p.ej., alguna componente separada más de 3σ de su media local), o bien emplear modificaciones del coste cuadrático que reduzcan dicha sensibilidad; así

- coste de Huber
  (cuadrático + lineal)



- coste de Talvar
  (cuadrático + constante)



(e_0 se elige en función de una medida robusta de la dispersión de e).

Ejercicio de Ampliación

A: ***(Principio de Invariancia)***

Determínese el estimador $\hat{s}_C$ *que corresponde a la aplicación de un coste para C(e) = C(-e) a la estimación de una va s condicionalmente gaussiana (s|* ***x*** *es gaussiana).*

$$\hat{s}_c : \min_{\hat{s}} \int_{-\infty}^{\infty} C(s-\hat{s}) p(s \mid \mathbf{x}) ds$$

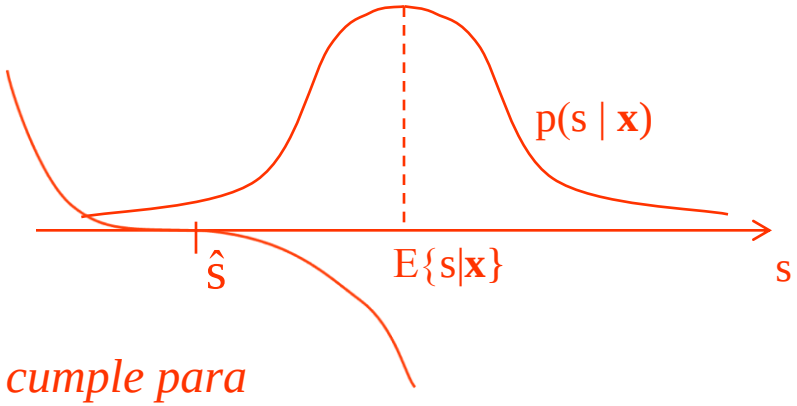
$\dot{C}$ *es impar: admitiendo (como es típico) que C(0) = 0 y que crece con |e|, es*

- *negativa para* $e = s - \hat{s} < 0$
- *positiva para* $e = s - \hat{s} > 0$

con lo que la minimización conduce a

$$\int_{\hat{s}_c}^{\infty} \dot{C}(s-\hat{s}_c) p(s \mid \mathbf{x}) ds = -\int_{-\infty}^{\hat{s}_c} \dot{C}(s-\hat{s}_c) p(s \mid \mathbf{x}) ds$$

gráficamente se tiene



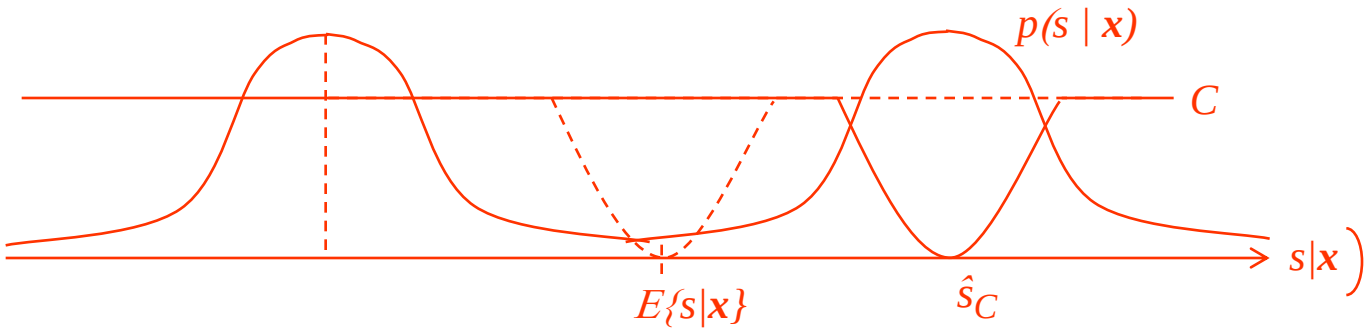
y la igualdad se cumple para

$$\hat{s}_C = E\{s \mid \boldsymbol{x}\} = \hat{s}_{ms}$$

Lo mismo podría haberse hecho buscando directamente el mínimo del coste medio,

$$\int_{-\infty}^{\infty} C(s-\hat{s})\, p(s \mid \boldsymbol{x})\, ds$$

– *Nótese que, si $\dot{C}$ es estrictamente creciente (C estrictamente cóncava), la conclusión es válida para cualquier p(s| **x**) que sea simétrica respecto a E{s| **x**}: ésta es la forma general del **Principio de Invariancia** para ddp a posteriori simétricas: si C(e) es par y estrictamente cóncava y p(s| **x**) es simétrica respecto a su media, $\hat{s}_C = \hat{s}_{ms}$ (si no hay concavidad estricta, no ha de ser así:*



– *Nótese que se incluyen todas las formas de la norma p (>1) de Minkowski, y sus versiones sin radicación, $|e|^p$.*

– *Nótese que lo anterior manifiesta una importante característica de robustez de $\hat{s}_{ms}$.*

Principio de Invariancia para $p(s \mid \mathbf{x})$ arbitraria

$$\hat{s}_C = \hat{s}_{ms} \quad \text{si y sólo si} \quad \frac{\partial C(s,\hat{s})}{\partial \hat{s}} = g(\hat{s})(s-\hat{s}) \qquad \text{con } g(\hat{s}) < 0$$

(v. Apéndice)

- Queda incluido el coste cuadrático: $g(\hat{s}) = -2$

- Hay otros muchos que lo verifican; p.ej., si s y $\hat{s}$ se mantienen entre 0 y 1, el coste de Hopfield-Hinton

$$C_{HH}(s,\hat{s}) = -s \ln \hat{s} - (1-s)\ln(1-\hat{s})$$

$$\left(\frac{\partial C_{HH}}{\partial \hat{s}} = -\frac{s}{\hat{s}} + \frac{1-s}{1-\hat{s}} = \frac{1}{\hat{s}(1-\hat{s})}\left[(1-s)\hat{s} - s(1-\hat{s})\right] = -\frac{1}{\hat{s}(1-\hat{s})}(s-\hat{s}) \right)$$

Si se emplean costes de este tipo para diseños máquina $\hat{s}_{\mathbf{w}}(\mathbf{x})$ y se aplica el algoritmo de gradiente

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \eta g\left[\hat{s}_{\mathbf{w}}\left(\mathbf{x}^{(k)}\right)\right]\left[s^{(k)} - \hat{s}_{\mathbf{w}}\left(\mathbf{x}^{(k)}\right)\right]\left.\frac{\partial \hat{s}_{\mathbf{w}}(\mathbf{x})}{\partial \mathbf{w}}\right|_{\mathbf{x}=\mathbf{x}^{(k)}}$$

es obvio que g influye:

- en la velocidad de convergencia;
- en la solución final: donde $\left|g\left[\hat{s}_{\mathbf{w}}\left(\mathbf{x}^{(k)}\right)\right]\right|$ sea menor, se tolerará mayor error.

(es decir: influye del mismo modo que el término

$$\left.\frac{\partial \hat{s}_{\mathbf{w}}(\mathbf{x})}{\partial \mathbf{w}}\right|_{\mathbf{x}=\mathbf{x}^{(k)}}$$

).

Discusión

Los Principios de Invariancia establecen que los diseños analíticos son idénticos: pero no que lo sean los diseños máquina:

- primero, porque la arquitectura de la máquina no tiene por qué corresponder a la óptima,

- segundo, y aunque la arquitectura incluya o pueda aproximar ilimitadamente la óptima, los diseños resultarán distintos por efectos del muestreo de las distribuciones: si bien es cierto que estos efectos decrecen al crecer el número de muestras (representativas).

(Naturalmente, en lo anterior no se tienen en cuenta los efectos de mínimos locales).

B. Costes para decisión

(Supondremos, para sencillez en la discusión, que se busca minimizar la tasa de error).

En aproximaciones máquina, es típico emplear un coste que permita un tratamiento analítico, como el cuadrático

– en el caso binario, con una salida, o, de la máquina, y

$$\min_{\mathbf{w}} (d-o)^2 \quad , \text{ con } d = \pm 1 \quad \text{ó} \quad 0,1$$

– en el C-ario, con C salidas y

$$\min_{\mathbf{w}} \sum_{c=1}^{C} (i_c - o_c)^2 \quad , \text{con } i_c = \begin{cases} 1, & \text{si c es la decisión correcta} \\ -1 \text{ ó } 0, & \text{en otro caso} \end{cases}$$

Así, pueden resolverse los diseños en forma iterativa (ocasionalmente, bloque).

Pero debe resaltarse que no se está minimizando la tasa de errores, sino algo "parecido": haciendo $(d-o)^2$ $(o\,(i_c-o_c)^2)$ pequeño, habrá pocos errores.

La dificultad radica en que habría que conseguir

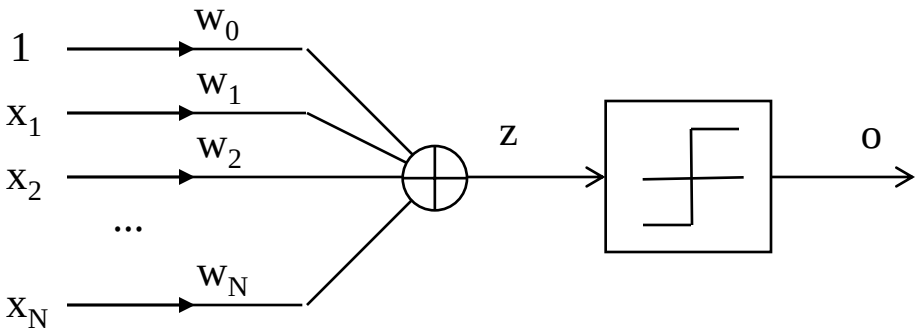
$$\min_{\mathbf{w}} \left| d - o_q \right|$$

siendo o_q también ± 1 ó 0, 1: y no existen procedimientos analíticos que permitan hacerlo en situaciones generales (análogamente, en C-arios).

Sí hay aproximaciones empíricas.

La Regla del Perceptrón

Supongamos una máquina con estructura lineal y cuantificación (± 1) a la salida (**Perceptrón Monocapa** "duro"):



que evidentemente define como frontera el hiperplano $\mathbf{w}_e^T \mathbf{x}_e = 0$

Una forma de entrenarlo es aplicar la **Regla del Perceptrón**: en versión secuencial

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \frac{\alpha}{2}\left[d^{(k)} - o^{(k)}\right]\mathbf{x}^{(k)} \quad (0 < \alpha < 1)$$

- no hay cambio si no hay error
- si $d^{(k)} = 1$, $o^{(k)} = -1$: $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \alpha\,\mathbf{x}^{(k)}$
- si $d^{(k)} = -1$, $o^{(k)} = 1$: $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \alpha\,\mathbf{x}^{(k)}$

así que $\mathbf{w}^{(k+1)T}\mathbf{x}^{(k)}$ cambia $\alpha \left\|\mathbf{x}^{(k)}\right\|_2^2$ en el sentido de corregir el error.

Se trata de supervisión por refuerzo Hebbiano (de nuevo), en lugar de supervisión por objetivo (minimización de un coste)

El refuerzo que se aplica es negativo: corrige los errores (positivo sería premiar los aciertos).

Nótese también que es un entrenamiento no lineal; intenta anular $d - o(\mathbf{w}) = d - \text{sgn}(\mathbf{w}^T\mathbf{x})$, que es función no lineal de los parámetros o pesos $\mathbf{w}$.

Trabajo: técnicas de aprendizaje por refuerzo.

Un primer inconveniente de lo propuesto es que el algoritmo sólo converge, en un número finito de pasos, si el problema es linealmente separable: si no, no se detiene.

Variantes heurísticas para enfrentarse a esta dificultad son:

- el **Algoritmo del Bolsillo**: conserva en una memoria aparte (bolsillo) los pesos que han dado lugar a la secuencia de pasos de entrenamiento libre de errores más larga: se acaba tras un alto número de pasos, y se adopta como solución el contenido del bolsillo;

- el **Algoritmo del Bolsillo con Trinquete** es análogo al anterior, pero sólo sustituye el contenido del bolsillo por un nuevo candidato tras verificar que éste da menos errores que el almacenado medidos sobre la totalidad de las muestras de entrenamiento.

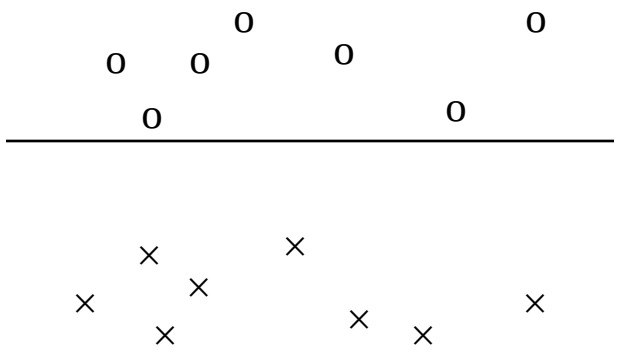
Si se extiende el principio de la Regla del Perceptrón a máquinas más generales, se presenta el mismo inconveniente: hay posibilidad de convergencia si el poder expresivo de la máquina permite una frontera de separación sin errores; caso contrario, el entrenamiento no se detiene.

Estas máquinas se llaman "**Basadas en Decisión**": en muchos casos conviene que el refuerzo sea analíticamente manejable, y así se corrige ante error mediante el gradiente de un coste C(d, z). En caso de salidas múltiples, suele dar buenos resultados el llamado **entrenamiento discriminativo**, relacionado con lo dicho, en el que se corrige positivamente la salida de la clase correcta y negativamente la mayor de las otras, siguiendo el **Principio de Mínima Perturbación** (no cambiar valores innecesariamente); si se hace esto sólo ante error, se tiene una versión generalizada de lo anterior.

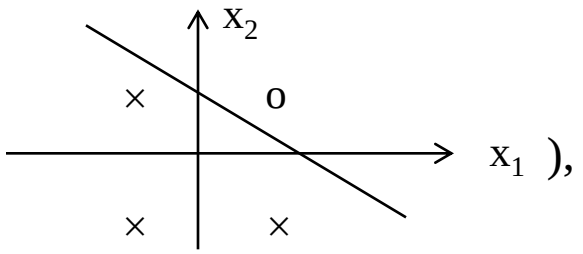
Otro inconveniente de la Regla del Perceptrón (y sus variantes) es que no ofrece buenas prestaciones en generalización:

- primero, porque, en todo caso, se consigue un número reducido de errores sobre el conjunto de entrenamiento, lo que no supone buena generalización;

- segundo, porque aunque el problema sea intrínsecamente separable (para la máquina usada), el algoritmo puede colocar la frontera en una posición poco adecuada

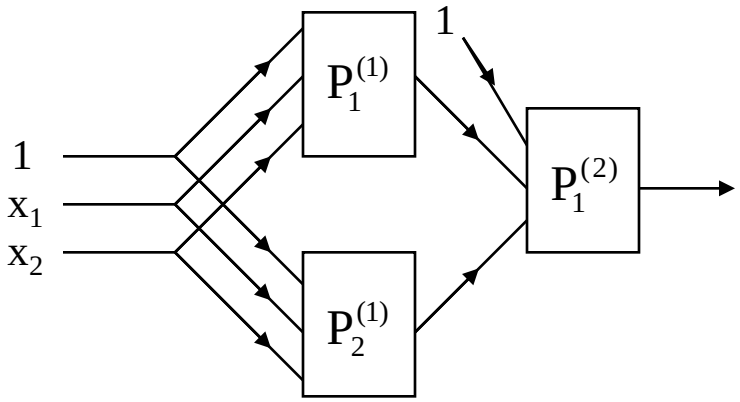


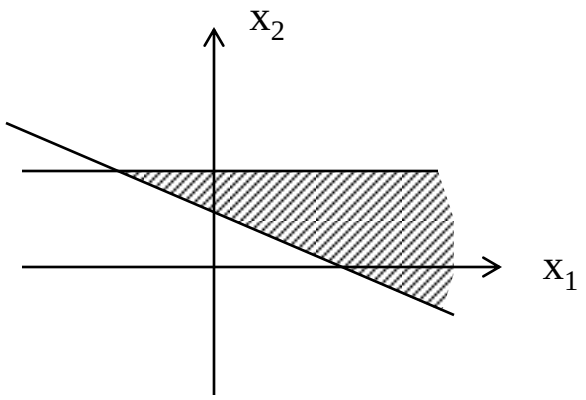
En principio, Perceptrones Monocapa "duros" pueden asociarse en capas consecutivas para ofrecer regiones de decisión más generales: así, si en el esquema que sigue $P_1^{(2)}$ realiza una función "AND" (lo que es inmediato:



),

la asociación daría regiones de decisión como se muestra a la derecha:





pero es fácil comprender que no puede extenderse el entrenamiento a los pesos de los perceptrones de la entrada (en general, interiores y de entrada).

El ADALINE ("Adaptive Linear Element") ("Adaptive Linear Neuron")

Tiene la misma arquitectura del Perceptrón Monocapa "duro", pero se entrena mediante el algoritmo de Widrow-Hoff aplicado al error a la entrada del decisor, $(d - z)^2$:

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \eta\left(d^{(k)} - z^{(k)}\right)\mathbf{x}^{(k)}$$

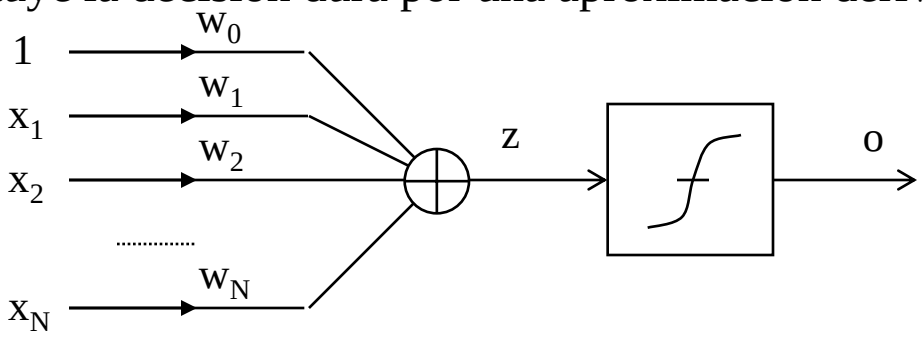
con lo que

- no se minimiza el número de errores; pero
- la convergencia está garantizada (a un mínimo local, en general) eligiendo adecuadamente η;
- las condiciones de generalización mejoran: la frontera tiende a situarse "equiespaciada" entre las nubes de muestras.

Subsiste la imposibilidad de entrenar estructuras multicapa.

La activación blanda

Se sustituye la decisión dura por una aproximación derivable



con lo que se puede aplicar gradiente sobre el valor cuadrático de

e = d – o = d – f(z) (que no es el número de errores)

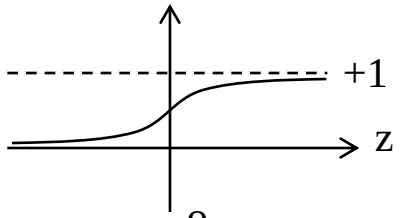
$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \frac{\eta}{2}\frac{\partial\left[d - f(z)\right]^2}{\partial \mathbf{w}} = \mathbf{w}^{(k)} + \eta\left[d - f(z)\right]\frac{\partial f(z)}{\partial \mathbf{w}} =$$

$$= \mathbf{w}^{(k)} + \eta\left[d - f(z)\right]\dot{f}(z)\frac{\partial \mathbf{w}^T\mathbf{x}}{\partial \mathbf{w}} = \mathbf{w}^{(k)} + \eta\left[d - f(z)\right]\dot{f}(z)\mathbf{x}$$

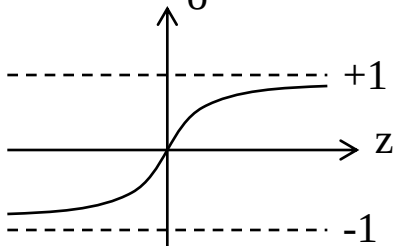
Nótese que, dada la forma de f(z), $\dot{f}(z) \to 0 \text{ con } z \to \pm\infty$: deteniendo la corrección; esta situación de parálisis puede evitarse fijando objetivos finales reducidos para o (p.ej,. $\pm$0.95 ó 0.95 y 0.05)

Es típico tomar:

$$(d = 1,0) \;:\; f(z) = sgm(z) = \frac{1}{1+e^{-z}}$$



$$(d = \pm 1) \;:\; f(z) = th(z) = \frac{1-e^{-z}}{1+e^{-z}}$$



(Puede usarse gz en lugar de z, con una ganancia g: pero ha de notarse que se trata simplemente de un factor de escala para las **w**, ya que $z = \mathbf{w}_e^T \mathbf{x}_e$)

– porque $\dot{f}$ tiene expresiones inmediatas en función de f = o, la propia salida de la no linealidad:

$$\dot{sgm} = o(1-o)$$

$$\dot{th} = 1-o^2$$

– porque si las hipótesis consideradas son gaussianas con iguales matrices de covarianza: H_i: G($\mathbf{m}_i$, V); resulta ser

$$\Pr(H_1 \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid H_1)\Pr(H_1)}{p(\mathbf{x})} = \frac{\exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T \mathbf{V}^{-1}(\mathbf{x}-\mathbf{m}_1)\right]\Pr(H_1)}{\sum_{i=1}^{2}\exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \mathbf{V}^{-1}(\mathbf{x}-\mathbf{m}_i)\right]\Pr(H_i)} =$$

$$= \frac{1}{1+\exp\left[-(\mathbf{m}_1-\mathbf{m}_0)^T \mathbf{V}^{-1}\mathbf{x} - \frac{1}{2}\mathbf{m}_1{}^T \mathbf{V}^{-1}\mathbf{m}_1 + \frac{1}{2}\mathbf{m}_0{}^T \mathbf{V}^{-1}\mathbf{m}_0 + \ln\frac{\Pr(H_0)}{\Pr(H_1)}\right]}$$

que tiene la forma $\frac{1}{1+\exp\left(-\mathbf{w}_e^T \mathbf{x}_e\right)}$: de modo que el decisor tiene

<u>capacidad expresiva</u> para estimar $\Pr(H_1 \mid \mathbf{x})$, lo que es una ventaja, según se verá de inmediato.

(Nótese que, si las matrices de covarianza fuesen diferentes, haría falta una <u>forma cuadrática</u> en el exponente)

– porque, quedando claro que no hay inconveniente ya para entrenar Perceptrones Multicapa, es posible demostrar que estas estructuras con una capa intermedia y empleando este tipo de activaciones son aproximadores universales.

(Del entrenamiento de los Perceptrones Multicapa, que ya son Redes Neuronales tradicionales, se dará detalle en otra Unidad)

En el caso de problemas C-arios, se seguirían disponiendo, como se indicó al principio de la discusión de los costes para decisión, de C salidas o_c, a cada una de las cuales llegaría una combinación lineal de las entradas; y, para el mismo fin de tener capacidad expresiva para representar las $Pr(H_c|\mathbf{x})$ en caso de hipótesis gaussiana de iguales covarianzas, se utilizará la activación softmax o de Potts

$$o_c = \frac{\exp(z_c)}{\sum_{i=1}^{C} \exp(z_i)}$$

Estimación de las probabilidades "a posteriori" (caso binario)

Con una máquina, si se emplea un objetivo C(d,o) tal que

$$\frac{\partial C(d,o)}{\partial o} = g(o)(d-o), \quad \text{con } g(o) < 0$$

la salida estimará $E\{d \mid \mathbf{x}\}$, según ya se sabe:

- si d=1,0 : $E\{d \mid \mathbf{x}\} = 1\ Pr(H_1 \mid \mathbf{x}) + 0\ Pr(H_0 \mid \mathbf{x}) = Pr(H_1 \mid \mathbf{x})$
- si d= ±1 : $E\{d \mid \mathbf{x}\} = 1\ Pr(H_1 \mid \mathbf{x}) - 1\ Pr(H_0 \mid \mathbf{x}) = 2\ Pr(H_1 \mid \mathbf{x}) - 1$

y lo hará

- según la representatividad de las observaciones
- de acuerdo con la capacidad expresiva de la máquina
- según el coste C aplicado, del modo que ya se ha discutido en estimación

En todo caso, conviene destacar que, al ser las muestras cercanas a la frontera pocas en problemas de decisión típicos (las ddp de las clases no estarán significativamente solapadas), la estimación tenderá a ser mala precisamente en su entorno; siendo allí donde más interés tiene (el verdadero problema es la decisión, que se toma en función de que "o" supere o no un umbral). Por ello, puede convenir:

- elegir una g que pondere especialmente las cercanías de la frontera ($o \cong 1/2$ con d = 1, 0 e iguales probabilidades a priori, p. ej)

  (lo que no implica necesariamente mejor decisión: además de los efectos de la limitación expresiva de la máquina, en muchos casos el coste HH, $-d \ln o - (1-d) \ln (1-o)$, da mejor resultado que el cuadrático, teniendo factores |g| iguales a $1/o(1-o)$ (que es mínimo para $o = 1/2$) y 2, respectivamente; la ventaja del HH parece radicar en que rechaza fuertemente salidas muy erróneas, ya que el correspondiente ln toma valores muy negativos);

– seleccionar para el entrenamiento muestras en un entorno de la frontera (a costa de empeorar la estimación en otras regiones)

(No deben omitirse en el entrenamiento las otras muestras: contribuyen mucho a un primer ajuste "grueso" de la frontera, lo que ayuda a una mejor selección – ya que las muestras a seleccionar son las próximas a la frontera, y, para proceder, ha de tenerse una idea de cuál es ésta – ; además, caso de querer implementar soluciones adaptativas, son las muestras lejanas las que permiten una rápida adaptación).

(Una mayor sutileza llevaría a considerar especialmente las muestras "críticas": aquellas que verdaderamente definen la frontera – p.ej. de dos muestras próximas a la frontera y muy próximas entre sí, sólo una resultaría crítica –; lamentablemente, se trata de un problema auto-referente, para el que sólo existen soluciones aproximadas).

Finalmente, nótese que el coste HH puede reescribirse

$$d \ln \frac{d}{o} + (1-d) \ln \frac{1-d}{1-o}$$

que es una forma simetrizada de la **distancia de Kullback-Leibler** entre distribuciones discretas de probabilidad, $\{P_l\}$ y $\{Q_l\}$

$$\sum_l P_l \ln \frac{P_l}{Q_l}$$

con lo que la minimización del coste HH puede considerarse como la búsqueda de las probabilidades "a posteriori" que (muestralmente) sean más cercanas a las distribuciones "degeneradas" dadas por los indicadores (1, 0) de la clase para las muestras de entrenamiento.

Trabajo: ¿Sería provechoso emplear otras distribuciones en lugar de la degenerada?

Trabajo: Otras medidas de semejanza entre distribuciones de probabilidad.

Apéndice

Prueba del Principio de Invariancia para ddp "a posteriori" arbitraria

$$\frac{\partial C(s,\hat{s})}{\partial \hat{s}} = g(\hat{s})(s-\hat{s}), \quad g(\hat{s}) < 0 \quad \Leftrightarrow \quad \hat{s}_C = \hat{s}_{ms}$$

a) Suficiente ($\Rightarrow$)

$$\left.\frac{\partial \overline{C}(\hat{s}\,|\,\mathbf{x})}{\partial \hat{s}}\right|_{\hat{s}=\hat{s}_C} = \left.\int_{-\infty}^{\infty} g(\hat{s})(s-\hat{s})p(s\,|\,\mathbf{x})ds\right|_{\hat{s}=\hat{s}_C} = 0 \quad \Rightarrow \quad \hat{s}_c = \int_{-\infty}^{\infty} sp(s\,|\,\mathbf{x})ds = \hat{s}_{ms}$$

$$\left.\frac{\partial^2 \overline{C}(\hat{s}\,|\,\mathbf{x})}{\partial \hat{s}^2}\right|_{\hat{s}=\hat{s}_C} = \left.-g(\hat{s})\int_{-\infty}^{\infty} p(s\,|\,\mathbf{x})ds + \dot{g}(\hat{s})\int_{-\infty}^{\infty}(s-\hat{s})p(s\,|\,\mathbf{x})ds\right|_{\hat{s}=\hat{s}_C} = -g(\hat{s}_C) > 0$$

a) Necesaria: admítase la forma general

$$\frac{\partial C(s,\hat{s})}{\partial \hat{s}} = g(\hat{s},s)(s-\hat{s})$$

entonces

$$\left.\frac{\partial \overline{C}(\hat{s}\,|\,\mathbf{x})}{\partial \hat{s}}\right|_{\hat{s}=\hat{s}_{ms}} = \int_{-\infty}^{\infty} g(\hat{s}_{ms},s)(s-\hat{s}_{ms})p(s\,|\,\mathbf{x})ds = 0$$

tendría que cumplirse para cualquier p(s|**x**);

en particular, para

$$p(s \mid \mathbf{x}) = \begin{cases} P, & s = s_1 \\ 1-P, & s = s_2 \end{cases}; \quad \text{que tiene:} \quad \hat{s}_{ms} = Ps_1 + (1-P)s_2$$

llevando la ddp a la condición anterior proporciona

$$\int_{-\infty}^{\infty} g(\hat{s}_{ms}, s)(s - \hat{s}_{ms})\left[P\delta(s - s_1) + (1-P)\,\delta(s - s_2)\right] ds = 0$$

$$Pg(\hat{s}_{ms}, s_1)(s_1 - \hat{s}_{ms}) + (1-P)g(\hat{s}_{ms}, s_2)(s_2 - \hat{s}_{ms}) = 0$$

y sustituyendo $\hat{s}_{ms}$ en las diferencias

$$Pg(\hat{s}_{ms}, s_1)\left[s_1 - Ps_1 - (1-P)s_2\right] + (1-P)g(\hat{s}_{ms}, s_2)\left[s_2 - Ps_1 - (1-P)s_2\right] = 0$$

$$Pg(\hat{s}_{ms}, s_1)(1-P)(s_1 - s_2) + (1-P)g(\hat{s}_{ms}, s_2)P(-s_1 + s_2) = 0$$

$$P(1-P)(s_1 - s_2)\left[g(\hat{s}_{ms}, s_1) - g(\hat{s}_{ms}, s_2)\right] = 0$$

lo que requiere que $g(\hat{s}_{ms}, s_1) - g(\hat{s}_{ms}, s_2) = 0$

y, como s_1 y s_2 son arbitrarios, se necesitaría $g(\hat{s}_{ms}, s) = g(\hat{s}_{ms})$

forma de la condición