..

APRENDIZAJE SUPERVISADO: LEARNING

PROBABILIDAD 2023 - RICARDO FRAIMAN

1. Introducción

Reconocimiento de patrones (pattern recognition), clasificación, machine learning, aprendizaje automático (learning) se ocupa de predecir la naturaleza desconocida de una observación - una cantidad discreta como blanco o negro, cero o uno, enfermo o sano, que llamaremos **clase**.

"El aprendizaje automático es un campo enorme que se ha aplicado a diversos problemas. En su núcleo está el problema de clasificación (supervisada), donde el objetivo básico es predecir las respuestas a preguntas de sí o no (¿Este paciente médico sufrirá un derrame cerebral el próximo año?, o ¿Esta persona hará clic en en este anuncio en línea? o tal vez ¿Este criminal será arrestado por otro crimen ¿el próximo año?, o ¿Esta persona dijo "sí" a la pregunta del asistente de voz?. Los algoritmos de clasificación leen las direcciones en el correo cuando lo envía por correo postal, leen los cheques en el banco, procesan sus palabras habladas cuando habla en su asistente de voz, e incluso cancelan los ecos cuando habla a través de una línea telefónica. Se utilizan para recomendar productos en Internet, proporcionar devoluciones de búsqueda, decidir sobre créditos, sugerir correcciones ortográficas al texto escrito y predecir los niveles de contaminación durante incendios forestales."

Una observación es una colección de medidas numéricas que representamos por un vector d-dimensional x. Por ejemplo una imagen (que es una secuencia de bits, una por pixel), un vector de datos en el tiempo, un electrocardiograma, o una firma en un cheque convenientemente digitalizada.

La naturaleza desconocida de una observación se llama una clase. Se denota por y y toma valores en un conjunto finito $\{1, \dots, M\}$. En lo que sigue por simplicidad consideraremos el caso binario, $y \in \{0, 1\}$.

En machine learning consideraremos funciones

$$g(x) : \mathbb{R}^d \to \{0, 1\}$$

que representan nuestra predicción de y dado x, y llamaremos clasificadores. Nuestro clasificador cometerá un error cuando $y \neq g(x)$.

1.1. **El modelo de aprendizaje.** Introducimos ahora el modelo probabilístico con que trabajaremos.

Sea (X, Y) un vector aleatorio en $\mathbb{R}^d \times \{0, 1\}$. La distribución de (X, Y) describirá la "frecuencia" de encontrar pares particulares en la práctica.

La distribución del par (X, Y) se puede describir de varias formas distintas. En particular lo podemos hacer por el par (μ, η), donde μ es la distribución de X y η es la esperanza condicional de Y dado X.

Mas precisamente, si $A \subset \mathbb{R}^d$ es un boreliano,

$$\mu(A) = \mathbb{P}(X \in A),$$

y para $x \in \mathbb{R}^d$

$$\eta(x) = \mathbb{E}(Y|X = x) = \mathbb{P}(Y = 1|X = x).$$

Proposition 1.1. *El par (μ, η) determina la distribución de (X, Y).*

Proof. Dado $C \subset \mathbb{R}^d \times \{0, 1\}$, tenemos que

$$C = \left(C \cap \{\mathbb{R}^d \times \{0\}\}\right) \bigcup \left(C \cap \{\mathbb{R}^d \times \{1\}\}\right) =: (C_0 \times \{0\}) \cup (C_1 \times \{1\}).$$

$$\mathbb{P}((X, Y) \in C) = \mathbb{P}(X \in C_0, Y = 0) + \mathbb{P}(X \in C_1, Y = 1) =$$

$$\mathbb{E}\left(\mathcal{I}_{C_0}(X)\mathcal{I}_0(Y)\right) + \mathbb{E}\left(\mathcal{I}_{C_1}(X)\mathcal{I}_1(Y)\right) =$$

$$\mathbb{E}\left(\mathcal{I}_{C_0}(X)\mathbb{E}(\mathcal{I}_0(Y)|X)\right) + \mathbb{E}\left(\mathcal{I}_{C_1}(X)\mathbb{E}(\mathcal{I}_1(Y)|X)\right) =$$

$$\mathbb{E}\left((\mathcal{I}_{C_0}(X)(1 - \eta(X))\right) + \mathbb{E}\left(\mathcal{I}_{C_1}(X)\eta(X)\right)) =$$

$$\int_{C_0} (1 - \eta(x))\mu(x) + \int_{C_1} \eta(x)\mu(x).$$

□

Cualquier función $g : \mathbb{R}^d \to \{0, 1\}$ define un clasificador. Un error ocurre cuando $g(X) \neq Y$ y la probabilidad de error para el clasificador g está dada por

$$L(g) = P(g(X) \neq Y). \tag{1}$$

El mejor clasificador será g^* definido por

$$g^* = arg \min_{g:\mathbb{R}^d \to \{0,1\}} L(g) = arg \min_{g:\mathbb{R}^d \to \{0,1\}} \mathbb{P}(g(X) \neq Y). \tag{2}$$

1.2. **El clasificador de Bayes.** El clasificador de Bayes se define como

$$g^*(x) = \mathcal{I}_{\{\eta(x)>1/2\}}, \tag{3}$$

o sea vale 1 si $\eta(x) > 1/2$ y 0 sino.

Theorem 1.2. *Para toda función de decisión $g : \mathbb{R}^d \to \{0,1\}$ (boreliana) tenemos que*

$$\mathbb{P}(g^*(X) \neq Y) \leq \mathbb{P}(g(X) \neq Y). \tag{4}$$

Proof.

$$\mathbb{P}(g(X) \neq Y|X = x) = 1 - \mathbb{P}(g(X) = Y|X = x) =$$
$$1 - [\mathbb{P}(Y = 1, g(X) = 1|X = x) + \mathbb{P}(Y = 0, g(X) = 0|X = x)] =$$
$$1 - \left[\mathbb{E}(\mathcal{I}_1(Y)\mathcal{I}_{\{g(X)=1\}}|X = x) + \mathbb{E}(\mathcal{I}_0(Y)\mathcal{I}_{\{g(X)=0\}}|X = x)\right] =$$
$$1 - \left[\mathcal{I}_{\{g(X)=1\}}\mathbb{P}(Y = 1|X = x) + \mathcal{I}_{\{g(X)=0\}}\mathbb{P}(Y = 0|X = x\right] =$$
$$1 - \left[\mathcal{I}_{\{g(X)=1\}}\eta(x) + \mathcal{I}_{\{g(X)=0\}}(1 - \eta(x))\right].$$

Luego, tenemos que para todo x

$$\mathbb{P}(g(X) \neq Y|X = x) - \mathbb{P}(g^*(X) \neq Y|X = x) =$$
$$\eta(x)\left[\mathcal{I}_{\{g^*(X)=1\}} - \mathcal{I}_{\{g(X)=1\}}\right] + (1 - \eta(x))\left[I_{\{g^*(X)=0\}} - \mathcal{I}_{\{g(X)=0\}}\right] = (^*)$$
$$(2\eta(x) - 1)(\mathcal{I}_{\{g^*(X)=1\}} - \mathcal{I}_{\{g(X)=1\}} \geq 0,$$

por la definición de g^*. Finalmente obtenemos el resultado tomando esperanza respecto a X.

(* Usamos que $\mathcal{I}_{\{g^*(X)=0\}} - \mathcal{I}_{\{g(X)=0\}} = \mathcal{I}_{\{g(X)=1\}} - \mathcal{I}_{\{g^*(X)=1\}}$) □

g^* se llama la regla de Bayes y $L^* = \mathbb{P}(g^*(X) \neq Y)$ la probabilidad Bayes de error o riesgo de Bayes. L^* mide la dificultad del problema y no tiene porqué ser cero.

Remark. La demostración anterior muestra además que

(1) $L(g) = 1 - \mathbb{E}\left(\mathcal{I}_{\{g(X)=1\}}\eta(X) + \mathcal{I}_{\{g(X)=0\}}(1 - \eta(X))\right)$, y en particular

(2) $L^* = 1 - \mathbb{E}\left(\mathcal{I}_{\{\eta(X)>1/2\}}\eta(X) + \mathcal{I}_{\{\eta(X)\leq 1/2\}}(1 - \eta(X))\right) =$

$$\mathbb{E}(min(\eta(X), (1 - \eta(X)))) = \frac{1}{2} - \frac{1}{2}\mathbb{E}(|2\eta(X) - 1|),$$

ya que $min(u, 1 - u) = \frac{1}{2} - \frac{1}{2}|2u - 1|$ para $0 \leq u \leq 1$.

Remark. Es claro que g^* depende de la distribución del par (X, Y). Si dicha distribución es conocida podemos calcular g^*, pero en general la distribución es desconocida y por tanto g^* lo es.

1.3. **Caso con densidades.** Sean f_1 la densidad de $X|Y=1$, f_0 la densidad de $X|Y=0$ y $p=\mathbb{P}(Y=1)$. Luego la densidad de X es $f=pf_1+(1-p)f_0$.

Ejercicio.

(1) Probar que

$$\eta(x)=\frac{pf_1(x)}{pf_1(x)+(1-p)f_0(x)},(^*)$$

y deducir que $g^*(x)=1$ si $\frac{f_1(x)}{f_0(x)}>(1-p)/p$ y 0 en otro caso.

(*)Para ello verificar que η cumple $\mathbb{E}(Yh(X))=\mathbb{E}(h(X)\mathbb{E}(Y|X)))$ para toda h integrable.

(2) Probar que $L^*=\int min(pf_1(x),(1-p)f_0(x))dx$.

(3) Si $p=1/2$ probar que $L^*=1/2-1/4\int|f_1(x)-f_0(x)|dx$. Sugerencia: usar que $min(a,b)=b-max(b-a,0)$.

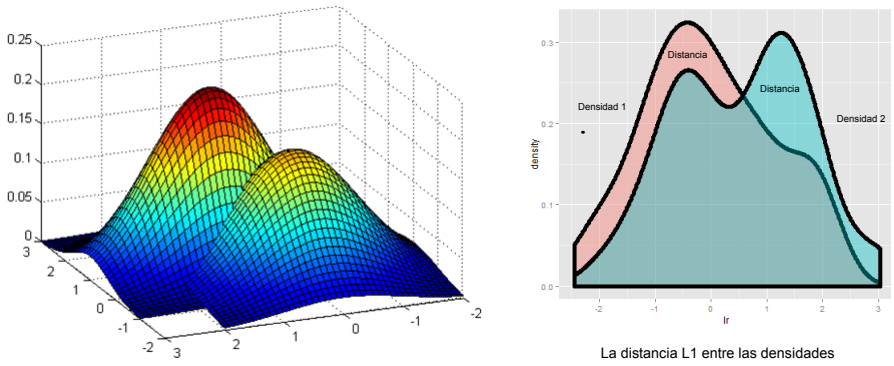

FIGURE 1.1. Caso con densidades.

2. Modelo no paramétrico: el caso realista

Si no conocemos la distribución del par (X,Y), supondremos que tenemos una muestra de entrenamiento

$$\aleph_n:=\{(X_1,Y_1),\dots(X_n,Y_n)\} \tag{5}$$

de vectores independientes e identicamente distrubuídos, con la misma distribución que (X,Y) donde los $Y's$ son observados.

Un clasificador se construye basado en la muestra de entrenamiento (5) a partir de una función g_n que predicirá el valor desconocido Y de una nueva observación X por

$$g_n(X,X_1,Y_1,\dots,X_n,Y_n),$$

donde X es el correspondiente a Y. El proceso de construcción de g_n se llama aprendizaje.

La idea básica subyacente es que a partir de $\{(X_1, Y_1), \dots (X_n, Y_n)\}$ podemos aproximar la distribución de (X, Y) y en particular considerar aproximaciones a η^* con la que construye la regla óptima de Bayes.

La performance de g_n se mide por la probabilidad condicional de error

(6)
$$L_n = L(g_n) = \mathbb{P}(g_n(X; X_1, Y_1, \dots, X_n, Y_n) \neq Y | X_1, Y_1, \dots, X_n, Y_n),$$

que es una variable aleatoria que depende de los datos $\aleph_n$.

Finalmente la calidad del clasificador la mediremos por

$$\mathbb{E}(L_n) = \mathbb{P}(g_n(X) \neq Y), \tag{7}$$

que a pesar de ser un promedio nos será muy útil si L_n se concentra alrededor de su media con probabilidad grande.

Definition 2.1. Una regla de clasificación es consistente si

$$\lim_{n\to\infty} \mathbb{E}(L_n) = L^*.$$

Si una regla de clasificación es consistente para toda posible distribución del par (X, Y) se dice que es universalmente consistente.

Remark. Como $L_n \leq 1$ (variables aleatorias uniformemente acotadas), entonces $\lim_{n\to\infty} \mathbb{E}(L_n) = L^*$ es equivalente a que L_n converja en probabilidad a L^*.

Hasta 1977 no se sabía si existían reglas universalmente consistentes (es decir sin hipótesis adicionales sobre la distribución del par (X, Y)). Todos los resultados ponían restricciones a dicha distribución. En 1977 Stone prueba un resultado muy general, y muestra que la regla de k-vecinos más cercanos (k-NN que analizaremos más adelante) es universalmente consistente si $k \to \infty$ y $k/n \to 0$.

La regla k-NN es muy simple de describir.

Dada la muestra de entrenamiento $\aleph_n$, procedemos como sigue:

(1) Ordenamos los vectores $(X_1, Y_1), \dots, (X_n, Y_n)$ de la muestra de entrenamiento de acuerdo a la distancia de X_i al x que queremos etiquetar.
$$\|X^{(1)} - x\| \leq \|X^{(2)} - x\| \leq \dots \leq \|X^{(n)} - x\|$$
(2) Supongamos además que no hay empates.

(3) Dado k impar definimos la etiqueta predicha y como la etiqueta más votada entre $Y^{(1)}, \dots Y^{(k)}$ correspondientes a los k vecinos más cercanos a x entre $X_1, \dots, X_n$.

O sea, consideramos la bola de centro x y radio (aleatorio) $H_n(x) = \|X^{(k)} - x\|$ votan los $Y^{(1)}, \dots Y^{(k)}$ correspondientes que están en la bola. Porqué debería funcionar? Como veremos más adelante, la condición $k/n \to 0$ garantizará que el radio $H_n(x)$ de la bola de centro x, tienda a cero, y la condición $k \to \infty$ que haya suficientes votantes en la bola.

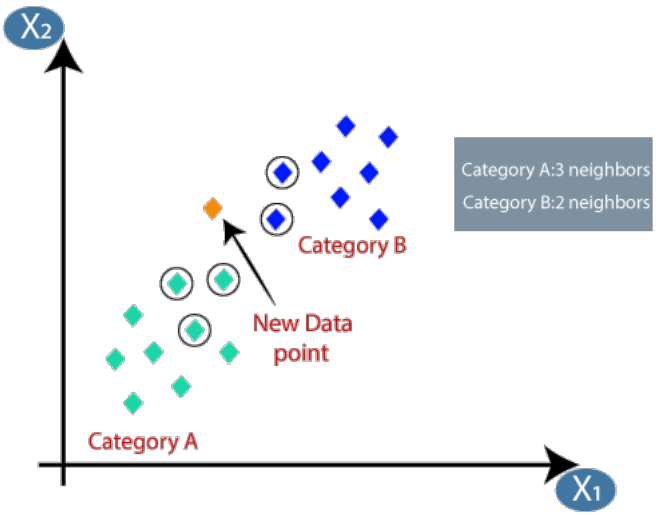

FIGURE 2.1. k-NN: vecinos más cercanos. k=5

3. Reglas de clasificación plug-in

Hemos visto que la mejor predicción de Y basada en la observación X es la dada por la regla de Bayes, que en el caso binario resulta ser

$$g^*(x) = 1, \text{ si } \eta(x) > 1 - \eta(x), \text{ y } 0 \text{ en otro caso.} \tag{8}$$

Luego, si conseguimos funciones $\widetilde{\eta}, 0 \leq \widetilde{\eta} \leq 1$ que aproximen a η, entonces resulta razonable considerar como regla de clasificación

$$g(x) = 1, \text{ si } \widetilde{\eta}(x) > 1 - \widetilde{\eta}(x), \text{ y } 0 \text{ en otro caso,} \tag{9}$$

o equivalentemente

$$g(x) = 1, \text{ si } \widetilde{\eta}(x) > 1/2 \text{ y } 0 \text{ en otro caso,} \tag{10}$$

para aproximar la regla de Bayes.

El siguiente teorema establece que si $\widetilde{\eta}$ está cerca de η (en L_1) entonces la probabilidad de error de g está cerca del de g^*.

Theorem 3.1. *Para la probabilidad de error de las reglas plug-in g tenemos que:*

(1) $\mathbb{P}(g(X) \neq Y) - L^* = 2\mathbb{E}(|\eta(X) - 1/2|\mathcal{I}_{\{g(X)\neq g^*(X)\}})$.
(2) $\mathbb{P}(g(X) \neq Y) - L^* \leq 2\mathbb{E}(|\eta(X) - \widetilde{\eta}(X)|)$.

Proof. Si para $x \in \mathbb{R}^d$, $g(x) = g^*(x)$, la diferencia entre las probabilidades condicionales de error es cero,

$$\mathbb{P}(g(X) \neq Y|X = x) - \mathbb{P}(g^*(X) \neq Y|X = x) = 0.$$

Sino, si $g(x) \neq g^*(x)$, de la demostración del Teorema 1.2, tenemos que

$$\mathbb{P}(g(X) \neq Y|X = x) - \mathbb{P}(g^*(X) \neq Y|X = x) =$$

$$(2\eta(x) - 1)\left(\mathcal{I}_{\{g^*(X)=1\}} - \mathcal{I}_{\{g(X)=1\}}\right) = |2\eta(x) - 1|\mathcal{I}_{\{g(x)\neq g^*(x)\}}.$$

Luego, tomando esperanza tenemos que

$$\mathbb{P}(g(X) \neq Y)) - L^* \leq \mathbb{E}(2|\eta(X) - 1/2|\mathcal{I}_{\{g(X)\neq g^*(X)\}}) \leq 2\mathbb{E}(|\eta(X) - \widetilde{\eta}(X)|),$$

ya que si $g(x) \neq g^*(x)$ entonces $|\eta(x) - \widetilde{\eta}(x)| \geq |\eta(x) - 1/2|$. (ya que si $g(x) \neq g^*(x)$ entonces $1/2$ está entre $\eta(x)$ y $\widetilde{\eta}(x)$ - o al revés). □

Remark. El método plug-in permitirá construir reglas de clasificación a partir de estimadores consistentes de la función de regresión no paramétrica $\eta(x)$, obteniendo reglas de clasificación consistentes como consecuencia del item (2) del teorema 3.1.

4. Otro enfoque: minimizando el riesgo empírico

Supongamos que tenemos una clase $\mathcal{C}$ de clasificadores $g : \mathbb{R}^d \to \{0.1\}$ y queremos encontrar uno en dicha clase con una pequeña probabilidad de error. Como no conocemos la distribución subyacente del par (X, Y), tendremos que usar nuevamente la muestra de entrenamiento para estimar las probabilidades de error de los clasificadores en la familia $\mathcal{C}$.

Un estimador natural de $L(g) = \mathbb{P}(g(X) \neq Y)$ es

$$\widehat{L}_n(g) = \frac{1}{n}\sum_{j=1}^{n} \mathcal{I}_{\{g(X_j)\neq Y_j\}}, \tag{11}$$

que llamaremos riesgo empírico de g. Un buen procedimiento será elegir un clasificador cuyo riesgo empírico sea mínimo en la clase $\mathcal{C}$. Si estimamos los errores de clasificación uniformemente bien en la clase $\mathcal{C}$, las cosas funcionarán bien. Si además la clase es "densa" en el espacio de clasificadores, funcionará muy bien !.

Proposition 4.1. *Sea g_n^* tal que*

$$\widehat{L}_n(g_n^*) \leq \widehat{L}_n(g) \quad \forall g \in \mathcal{C}.$$

Entonces $L(g_n^*) = \mathbb{P}(g_n^*(X) \neq Y|\aleph_n)$ verifica

(1) $L(g_n^*) - \inf_{g\in\mathcal{C}} L(g) \leq 2\sup_{g\in\mathcal{C}} |\widehat{L}_n(g) - L(g)|)$

(2) $|\widehat{L}_n(g_n^*) - L(g_n^*))| \leq \sup_{g\in\mathcal{C}} |\widehat{L}_n(g) - L(g)|$.

Proof. (1) $L(g_n^*) - \inf_{g\in\mathcal{C}} L(g) = L(g_n^*) - \widehat{L}_n(g_n^*) + \widehat{L}_n(g_n^*) - \inf_{g\in\mathcal{C}} L(g) \leq L(g_n^*) - \widehat{L}_n(g_n^*) + \sup_{g\in\mathcal{C}} |\widehat{L}_n(g) - L(g)| \leq 2\sup_{g\in\mathcal{C}} |\widehat{L}_n(g) - L(g)|$.

(2) es trivial. □

Remark. Por tanto cotas superiores para $\sup_{g\in\mathcal{C}} |\widehat{L}_n(g) - L(g)|$ nos da cotas superiores para dos cosas simultáneamente.

- (a) Una cota superior para la suboptimalidad de g_n^* en $\mathcal{C}$, o sea una cota para
$$L(g_n^*) - \inf_{g\in\mathcal{C}} L(g).$$
- (b) Una cota superior para el error $|\widehat{L}_n(g_n^*)) - L(g_n^*)|$ cuando usamos $\widehat{L}_n(g_n^*)$ para estimar la probabilidad de error $L(g_n^*)$ de la regla seleccionada.
- (c) Estudiar $\sup_{g\in\mathcal{C}} |\widehat{L}_n(g) - L(g)|$ corresponde a leyes de grandes números uniformes, correspondientes a resultados de procesos empíricos. En particular, la variable $n\widehat{L}_n(g)$ tiene distribución Binomial de parámetros $(n, L(g))$, luego para obtener cotas, hay que estudiar desviaciones uniformes de una binomial alrededor de su media. Para ello se utilizan resultados de desigualdades de concentración y la teoría de Vapnik-Chervonenkis, que exceden los objetivos de nuestro curso.

5. Teorema de Stone

El siguiente teorema nos permitirá deducir la consistencia universal de diversas reglas de clasificación, en particular de reglas plug-in.

Consideremos una regla de clasificación plug-in basada en un estimador de η de la forma

$$\eta_n(x) =: \sum_{i=1}^{n} \mathcal{I}_{\{Y_i=1\}} w_{ni}(x) = \sum_{i=1}^{n} Y_i w_{ni}(x), \tag{12}$$

donde los pesos $w_{ni}(x) = w_{ni}(x, X_1, \ldots, X_n)$ son no negativos y $\sum_{i=1}^{n} w_{ni}(x) = 1$.

Como regla de clasificación consideramos

$$g_n(x) = \mathcal{I}_{\{\eta_n(x))>1/2\}}. \tag{13}$$

Es intuitivamente claro que los pares (X_i, Y_i) tales que X_i está cerca de X proveerán mucho más información que aquellos que están lejos sobre η.

Luego, los pesos w_{ni} deben concentrar su masa alrededor de x y η_n se puede ver como un promedio local. Ejemplos de este tipo incluyen a las reglas basadas en histogramas, núcleos y vecinos más cercanos.

Theorem 5.1. *(Stone 1977). Supongamos que los pesos verifican las siguientes tres condiciones:*

(1) $\mathbb{E}(\sum_{i=1}^n w_{ni}(X) f(X_i)) \leq c\mathbb{E}(f(X))$.
(2) $\lim_{n\to\infty} \mathbb{E}(\sum_{i=1}^n w_{ni}(X) 1_{\|X_i - X\|>a}) = 0 \ \forall a > 0$.
(3) $\lim_{n\to\infty} \mathbb{E}(\max_{1\leq i\leq n} w_{ni}(X)) = 0$,

entonces g_n es universalmente consistente.

La condición (2) pide que los pesos se concentren alrededor de X. La condición (3) pide que no haya ningún peso que sea demasiado grande.

Proof. Supondremos además por simplicidad que la función de regresión η es continua de soporte compacto y por tanto uniformemente continua. **En este caso también la hipótesis (1) no es necesaria.** De acuerdo con el Teorema 1.2, basta probar que para toda distribución del par (X, Y),

$$\lim_{n\to\infty} \mathbb{E}((\eta(x) - \eta_n(X)^2)) = 0. \tag{14}$$

Denotaremos por $\widetilde{\eta}_n(x) = \sum_{i=1}^n \eta(X_i) w_{ni}(x)$ - que no es observable pues depende de η.

Luego, usando la desigualdad $(a+b)^2 \leq 2(a^2+b^2)$ tenemos que

$$\mathbb{E}((\eta(X) - \eta_n(X))^2) = \mathbb{E}((\eta(X) - \widetilde{\eta}_n(X) + \widetilde{\eta}_n(X) - \eta_n(X))^2) \leq 2(I + II),$$

donde $I = \mathbb{E}((\eta(X)) - \widetilde{\eta}_n(X))^2)$ e $II = \mathbb{E}((\widetilde{\eta}_n(X)) - \eta_n(X))^2)$, y por tanto bastará con acotar ambos términos.

$$I = \mathbb{E}((\eta(X)) - \widetilde{\eta}_n(X))^2) = \mathbb{E}(\left[\sum_{i=1}^n w_{ni}(X)(\eta(X) - \eta(X_i))\right]^2), \tag{15}$$

y como los pesos son no negativos y suman uno, son una probabilidad y podemos entonces acotar I usando el lado izquierdo de la igualdad por

$$\mathbb{E}(\sum_{i=1}^n w_{ni}(X)[\eta(X) - \eta(X_i)]^2),$$

usando la desigualdad de Jensen con $\phi(t) = t^2$. (Si ϕ es una función convexa, y X una variable aleatoria, $\phi(\mathbb{E}(X))) \leq \mathbb{E}(\phi(X)))$. Como η es uniformemente continua, dado $\epsilon > 0$ sea $a > 0$ tal que si $\|x_i - x\| \leq a$ entonces $|\eta(x_i) - \eta(x)| < \epsilon$. Luego como además $|\eta(x_i) - \eta(x)| \leq 1$ tenemos que

$$\mathbb{E}(\sum_{i=1}^{n} w_{ni}(X)[\eta(X)-\eta(X_i)]^2) \leq \mathbb{E}(\sum_{i=1}^{n} w_{ni}(X)1_{\|X-X_i\|\geq a})+\mathbb{E}(\sum_{i=1}^{n} w_{ni}(X)\epsilon) \to \epsilon,$$

por la hipótesis (2).

Resta acotar el término II.

Observemos que por la independencia

$$\mathbb{E}((Y_i - \eta(X_i))(Y_j - \eta(X_j))|X, X_1, \ldots, X_n)) = 0 \quad \forall i \neq j.$$

Luego,

$$\mathbb{E}([\widetilde{\eta}_n(X) - \eta_n(X)]^2) = \mathbb{E}([\sum_{i=1}^{n} w_{ni}(X)(\eta(X_i) - Y_i)]^2) =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}(w_{ni}(X)(\eta(X_i) - Y_i)w_{nj}(X)(\eta(X_j)) - Y_j)) =$$

$$\sum_{i=1}^{n} \mathbb{E}(w_{ni}(X)^2(\eta(X_i)) - Y_i)^2) \leq \mathbb{E}(\sum_{i=1}^{n} w_{ni}(X)^2) \leq$$

$$\mathbb{E}(\max_{1\leq i\leq n} w_{ni}(X) \sum_{i=1}^{n} w_{ni}(X)) = \mathbb{E}(\max_{1\leq i\leq n} w_{ni}(X)) \to 0,$$

por la condición (3). □

6. Aplicación a reglas k-NN

Veremos que si $k \to \infty$ y $k/n \to 0$ la regla k-NN es universalmente consistente.

Recordemos que la regla $g_n(x)$ actuaba por mayoría de votos entre las $Y's$ correspondientes a los k vecinos más cercanos a x. Sean $Y^{(1)}, \ldots, Y^{(k)}$ las etiquetas correspondientes a los $X's$ vecinos más cercanos a x.

Entonces $g_n(x)$ vale 1 si $\sum_{i=1}^{k} \mathcal{I}_{\{Y^{(i)}=1\}} > \sum_{i=1}^{k} \mathcal{I}_{\{Y^{(i)}=0\}}$, en que los empates se rompen al azar.

Theorem 6.1. *Si $k \to \infty$ y $k/n \to 0$, entonces para toda distribución $\mathbb{E}(L_n) = L^*$, siendo L_n el error correspondiente a la regla g_n.*

Proof. Basta probar que se cumplen las hipótesis del teorema de Stone. Los pesos $w_{ni}(X)$ valen $1/k$ is X_i está entre los k-vecinos más cercanos a X y cero sino. Como $k \to \infty$ se cumple la condición (3).

Si $\mathbb{P}(\|X^{(k)} - X)\| > a) \to 0$ - que vale si $k/n \to 0$, entonces $\mathbb{E}(\sum_{i=1}^{n} w_{ni}(X))\mathcal{I}_{\{\|X_i - X\| > a\}}) \to 0$ y por tanto vale la condición (2).

Para verificar la condición (1) en el caso general (ya observamos que si η es uniformemente continua no es necesaria) hay que probar que para toda f integrable

$$\mathbb{E}(\sum_{i=1}^{n} \frac{1}{k} 1_{\{X_i \text{está entre los k-vecinos mas cercanos a } X\}} f(X_i)) \leq c\mathbb{E}(f(X))$$

para alguna constante $c > 0$. No lo probaremos. Ver Lema 5.3 en Devroye-Gyorfi-Lugosi por ejemplo. □

7. Redes neuronales

Para el caso de las reglas basadas en hiperplanos (el perceptrón), el problema de minimizar el riesgo empírico en la familia

$$f(x, w) == \sigma\left(x_i^t w\right),$$

donde $\sigma(u) = 0$ si $u < 0$, $\sigma(u) = 1$ si $u \geq 0$, consiste en dada la muestra de entrenamiento $\aleph_n$ encontrar w que minimize

$$\widehat{L}_n(f) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i, w))^2. \tag{16}$$

En términos de redes neuronales el perceptrón corresponde a una red neuronal sin capas ocultas, que graficamente se representa como sigue.

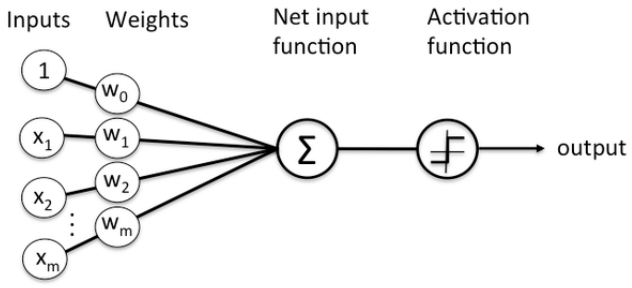

Figure 7.1. Red sin capas ocultas con una sola neurona.

No es posible usar el método basado en el gradiente para minimizar el riesgo empírico dado por (16) pues estamos usando la función σ que

toma valores 0 y 1. Ello motivó en el contexto de redes neuronales que veremos a continuación reemplazar la función σ por funciones mas regulares llamadas sigmoides, donde σ es una función continua, "suave" como por ejemplo

$$\sigma_1(u) = \frac{1}{1+e^{-u}}, \quad \sigma_2(u) = \frac{2arctg(u)+\pi}{2\pi}, \quad \sigma_{ReLU} = max(x,0),$$

entre otras.

Por otro lado sabemos que el perceptrón no es capaz de resolver muchos problemas salvo que los datos sean separables por un hiperplano. Para incrementar la flexibilidad del conjunto de reglas de decisión de aprendizaje consideraremos muchas neuronas (combinaciones lineales) y compondremos los outputs de cada neurona en una o varias capas ocultas (multilayer perceptrons).

Denotemos por $x = (x_1, \ldots, x_d)$ y

$$\psi(x) = c_0 + c^t x,$$

donde $x \in R^d$, $c \in R^d$ y $c_0 \in R$.

En una red neuronal con k neuronas y una capa oculta tomaremos

$$\psi(x) = c_0 + \sum_{i=1}^{k} c_i \sigma(\psi_i(x)),$$

donde los c_i son como antes y cada ψ_i es de la forma

$$\psi_i(x) = b_i + \sum_{i=1}^{d} a_{ij} x_j,$$

para ciertas constantes b_i y a_{ij}.

En la primera etapa calculamos k funciones ψ_i y le aplicamos la función sigmoide a cada una. En la siguiente etapa combinamos los outputs de $\sigma(\psi_i)$, $i = 1, \ldots, k$ y producimos con un threshold la salida $0, 1$.

En este caso diremos que tenemos k neuronas ocultas en que el output de la i–esima neurona es $u_i = \sigma(\psi_i(x)$, y $\psi(x) = c_0 + \sum_{i=1}^{k} c_i u_i$.

Podemos proseguir este proceso creando otras capas ocultas. Por ejemplo si tenemos dos capas ocultas

$$\psi(x) = c_0 + \sum_{i=1}^{l} c_i z_i,$$

con

$$z_i = \sigma(d_{i0} + \sum_{j=1}^{k} d_{ij} u_j) \quad 1 \leq i \leq l,$$

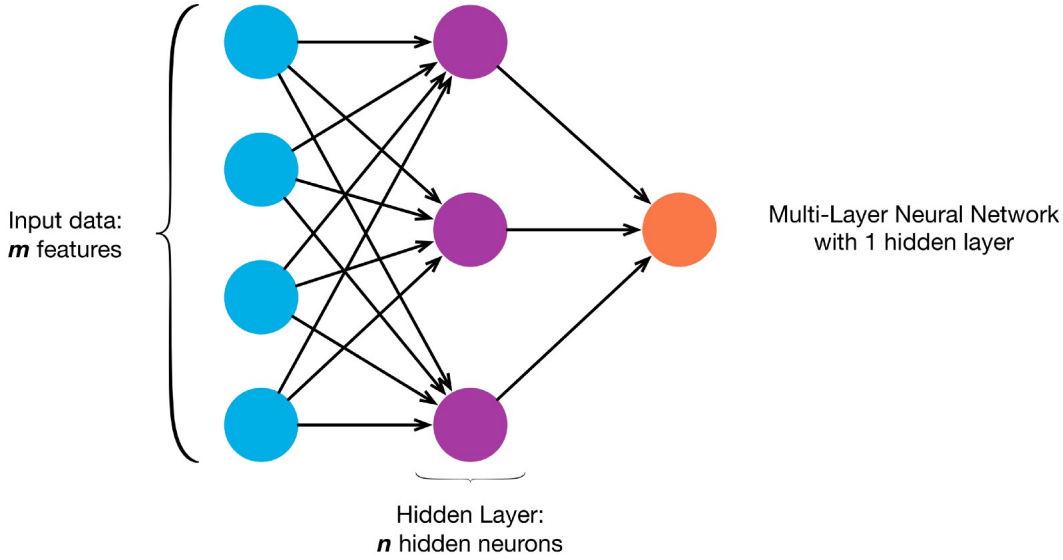

FIGURE 7.2. Red con una capa oculta

y

$$u_j = \sigma(b_j + \sum_{i=1}^{d} a_{ji} x_i) \quad 1 \leq j \leq k.$$

El primer layer oculto tiene k neuronas y el segundo layer oculto tiene l neuronas ocultas.

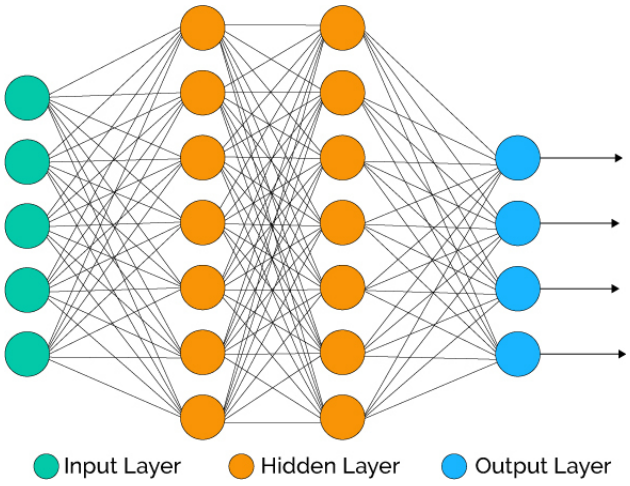

FIGURE 7.3. Red con dos capas ocultas

Pasar del perceptrón a una red con una capa oculta es no trivial. Sabemos que la discriminación lineal no es universalmente consistente, sin embargo una red con una capa oculta si lo es si permitimos que $k = k(n) \to \infty$. Por otro lado el problema de la minimización se complica significativamente y lo analizaremos más adelante.

8. Un primer resultado de consistencia

Analizaremos primero el caso general en que la clase de funciones $f \in \mathcal{F}$ de funciones se hace densa en la norma del supremo (L^∞) en el espacio de las funciones continuas $C[a,b]$ donde $[a,b]$ es un hiperrectángulo en R^d (una condición un poco fuerte). Omitiremos las demostraciones de estos resultados.

Proposition 8.1. *Supongamos que una sucesión de clases de funciones $\mathcal{F}_k$ se hace densa en la norma L^∞ en el espacio de las funciones continuas $C[a,b]$. Es decir, supongamos que $\forall a, b \in R^d$ y toda función acotada g*

$$\lim_{k\to\infty} \inf_{f\in\mathcal{F}_k} \sup_{x\in[a,b]} |f(x) - g(x)| = 0,$$

entonces para toda distribución de (X,Y),

$$\lim_{k\to\infty} \inf_{\Phi\in\mathcal{C}_k} L(\Phi) - L^* = 0,$$

donde $\mathcal{C}_k$ es la familia de los clasificadores de la forma

$$\Phi(x) = \mathcal{I}_{\{f(x)>\frac{1}{2}\}}, \quad f \in \mathcal{F}_k.$$

Los primeros resultados de consistencia para redes neuronales con una capa oculta son de (1989). Cybenko; Hornik, Stinchcombe y White; y Fanahasti que probaron independientemente que las redes neuronales (forward) con una capa oculta son densas con respecto a la norma del supremo en conjuntos acotados en el espacio de funciones continuas.

En otras palabras, tomando k suficientemente grande, toda función continua de R^d se puede aproximar arbitrariamente, uniformemente sobre conjuntos acotados por funciones obtenidas por redes neuronales con una capa oculta.

Theorem 8.2. *Para toda función continua $f : [a,b]^d \to R$, $[a,b]^d \subset R^d$ y para todo $\epsilon > 0$ existe una red neuronal con una capa oculta, cuya función asociada es*

$$\psi(x) = c_0 + \sum_{i=1}^{k} c_i \sigma(\psi_i(x)) \quad con \quad \psi_i(x) = b_i + \sum_{i=1}^{k} a_{ij} x_j,$$

que verifica

$$\sup_{x\in[a,b]^d} |f(x) - \psi(x)| < \epsilon.$$

Corollary 8.3. *Sea $\mathcal{C}_k$ una clase de clasificadores que contenga todas las redes neuronales con una capa oculta , k neuronas y sigmoide σ. Entonces para toda distribución del par (X, Y)*

$$\lim_{k\to\infty} \inf_{\psi\in\mathcal{C}_k} L(\psi) - L^* = 0.$$

9. Redes profundas: deep learning

Deep learning no es otra cosa que usar redes neuronales con muchas capas ocultas.

En cada capa llegan pesos y el output se obtiene aplicando una función no lineal ("activation function") $\sigma : R \to R$ a la suma ponderada de los inputs.

Si el input de la red es $x \in R^d$, la primer capa transforma a x en $f_1(x)$. La segunda capa transforma $f_1(x)$ en $f_2(f_1(x))$ y asi sucesivamente. El output de la k–esima capa oculta es la predicción final

$$f(x) = f_k(f_{k-1}(\dots(f_1(x))\dots).$$

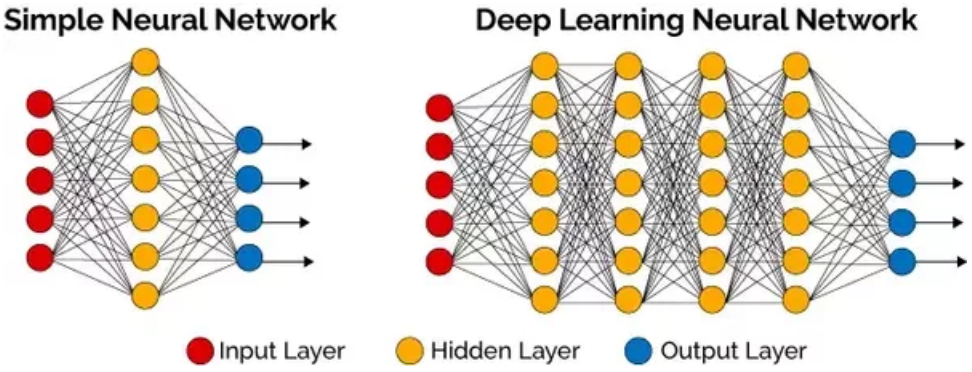

Figure 9.1. Izquierda: red con una capa oculta. Derecha: red con 4 capas ocultas.

Las redes se entrenan usando SGD, a veces llamado también back propagation. A los efectos de poder diferenciar se utilizan funciones σ suaves como

$$tgh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}; \quad \frac{1}{1+e^{-x}}; \quad max(x, 0),$$

entre otras, aunque últimamente se usa mucho $\sigma_{ReLU}(x) = max(x, 0)$ (rectified linear unit) que suele ser mas eficiente para redes con muchas capas ocultas.

Como las redes con una capa oculta ya son consistentes, lo serán las redes profundas. Entonces, al respecto hay dos problemas a analizar:

(1) Porqué redes con muchas capas ocultas son mejores que con pocas? En realidad la pregunta correcta es **cuando son mejores, es decir para qué problemas lo son?**
(2) Porqué el algoritmo SGD "funciona" a pesar de que la función a optimizar no es convexa?

Las respuestas a estas preguntas solo tienen resultados muy parciales. A pesar de ello, el aprendizaje profundo ha hecho avances muy significativos en resolver problemas que se habían resistido a los mejores intentos de la comunidad de inteligencia artificial por muchos años. En particular para:

- reconocimiento de imágenes
- reconocimiento de lenguaje hablado
- reconstrucción de circuitos cerebrales
- predicción de efectos de mutaciones en DNA
- traducción de lenguajes

entre otros muchos.

Para otros problemas, los métodos mas clásicos pueden funcionar mejor que las redes profundas, de modo que entender y poder responder la primer pregunta es muy importante.

Un argumento heurístico que se usa en el caso de imágenes, es que "los datos vienen dados en términos de "features" de bajo nivel, como valores de la intensidad en los pixeles. El objetivo es obtener alguna información de mas alto nivel, como que objetos hay en la imagen y que están haciendo. Las capas ocultas intentan transformar los features de bajo nivel en otros de alto nivel".