

Machine Learning

El problema del aprendizaje

Factibilidad del aprendizaje

Angel Vázquez-Patiño
angel.vazquezp@ucuenca.edu.ec

Departamento de Ciencias de la Computación
Universidad de Cuenca

14 de octubre de 2017

Objetivos

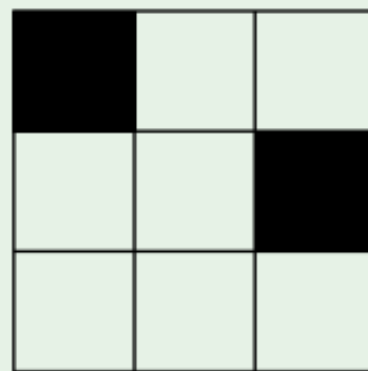
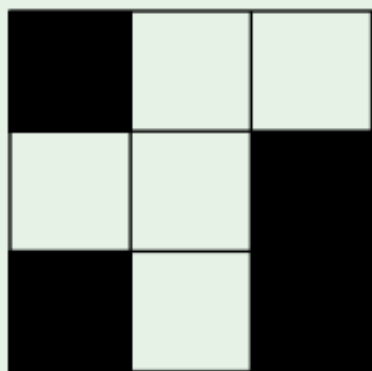
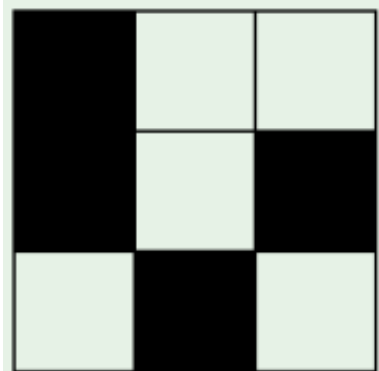
1. Entender la desigualdad de Hoeffding
2. Entender el trade-off complejidad del hypothesis set y error del modelo

Contenido

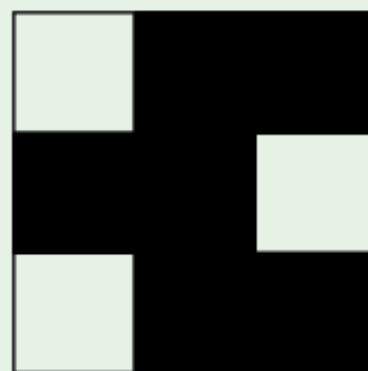
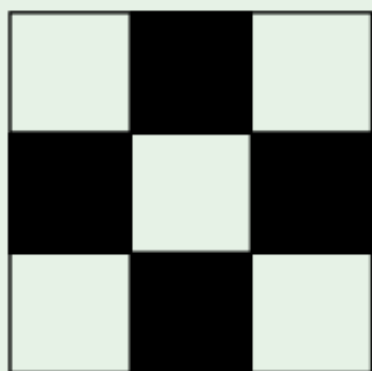
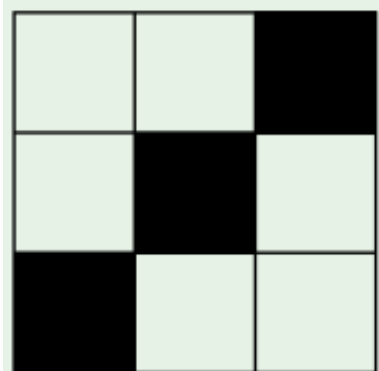
Más allá de los datos

Probabilidad al rescate

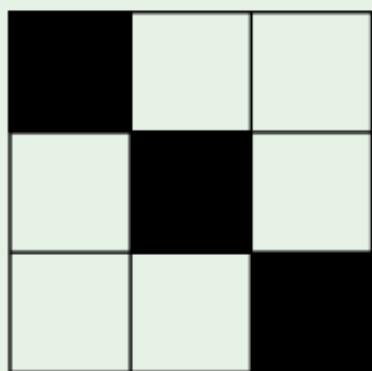
Factibilidad del aprendizaje



$$f = -1$$



$$f = +1$$



$$f = ?$$

Más allá de los datos

Más allá de los datos

- Cuando tenemos \mathcal{D} , conocemos el valor de f en todos los puntos de \mathcal{D}
- ¿Se aprendió algo de f ?
- Se sabe algo más allá de \mathcal{D}

Aprendizaje vs memorización

- Se sabe algo más allá \rightarrow aprendizaje
- Caso contrario no es factible
- Generalización

Más allá de los datos

- Se probará que f continua siendo desconocido más allá de los datos con un caso concreto

$$\mathcal{X} = \{0, 1\}^3$$

$\mathcal{D} \longrightarrow$

\mathbf{x}_n	y_n
0 0 0	○
0 0 1	●
0 1 0	●
0 1 1	○
1 0 0	●

Más allá de los datos

\mathbf{x}	y	g	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
0 0 0	○	○	○	○	○	○	○	○	○	○
0 0 1	●	●	●	●	●	●	●	●	●	●
0 1 0	●	●	●	●	●	●	●	●	●	●
0 1 1	○	○	○	○	○	○	○	○	○	○
1 0 0	●	●	●	●	●	●	●	●	●	●
1 0 1		?	○	○	○	○	●	●	●	●
1 1 0		?	○	○	●	●	○	○	●	●
1 1 1		?	○	●	○	●	○	●	○	●

Más allá de los datos

Exercise 1.7

For each of the following learning scenarios in the above problem, evaluate the performance of g on the three points in \mathcal{X} outside \mathcal{D} . To measure the performance, compute how many of the 8 possible target functions agree with g on all three points, on two of them, on one of them, and on none of them.

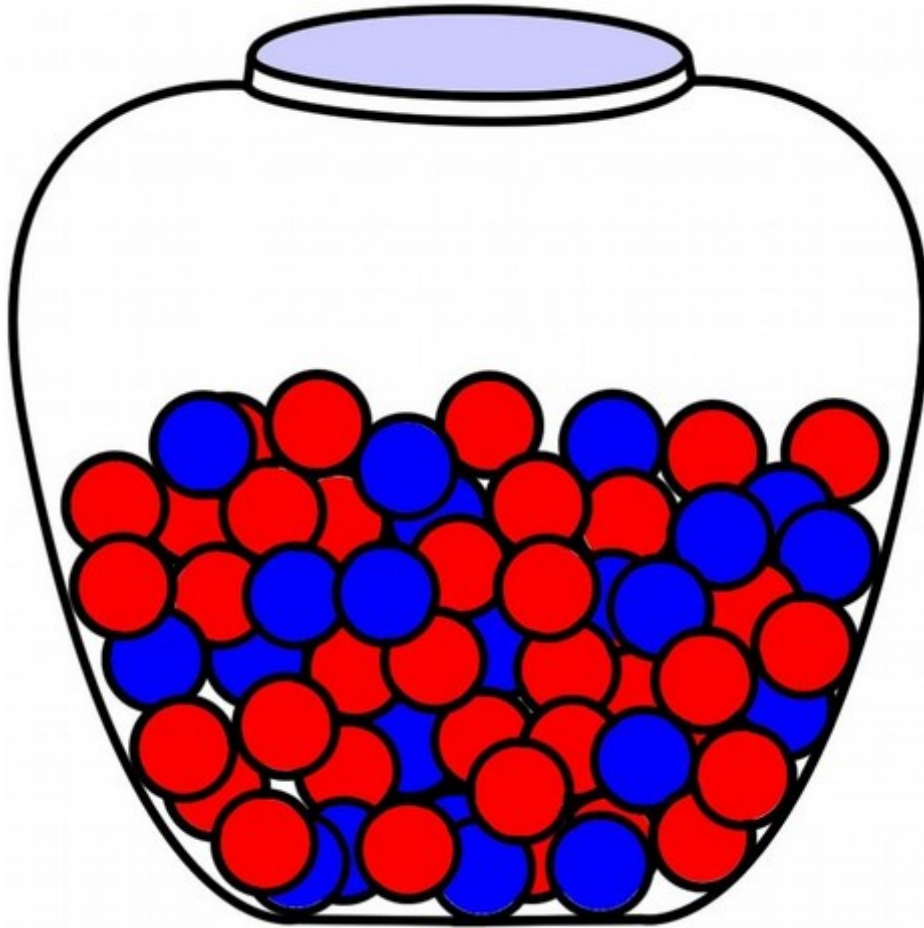
- (a) \mathcal{H} has only two hypotheses, one that always returns '●' and one that always returns '○'. The learning algorithm picks the hypothesis that matches the data set the most.
- (b) The same \mathcal{H} , but the learning algorithm now picks the hypothesis that matches the data set the *least*.
- (c) $\mathcal{H} = \{\text{XOR}\}$ (only one hypothesis which is always picked), where XOR is defined by $\text{XOR}(\mathbf{x}) = \bullet$ if the number of 1's in \mathbf{x} is odd and $\text{XOR}(\mathbf{x}) = \circ$ if the number is even.
- (d) \mathcal{H} contains all possible hypotheses (all Boolean functions on three variables), and the learning algorithm picks the hypothesis that agrees with all training examples, but otherwise disagrees the most with the XOR.

Probabilidad al rescate

Probabilidad al rescate

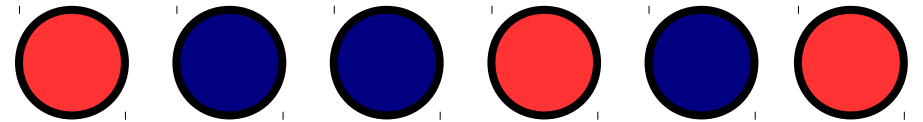
- Se puede inferir algo fuera de \mathcal{D} usando solamente \mathcal{D} pero en una forma probabilística
- Lo que se infiera puede no ser mucho comparado a aprender la f completa pero establecerá el principio que se puede conseguir fuera de \mathcal{D}
- Una vez establecido eso, se tomará eso para el problema de aprendizaje general y demostrará qué se puede y qué no se puede aprender

Probabilidad al rescate



- μ = probabilidad de canicas rojas

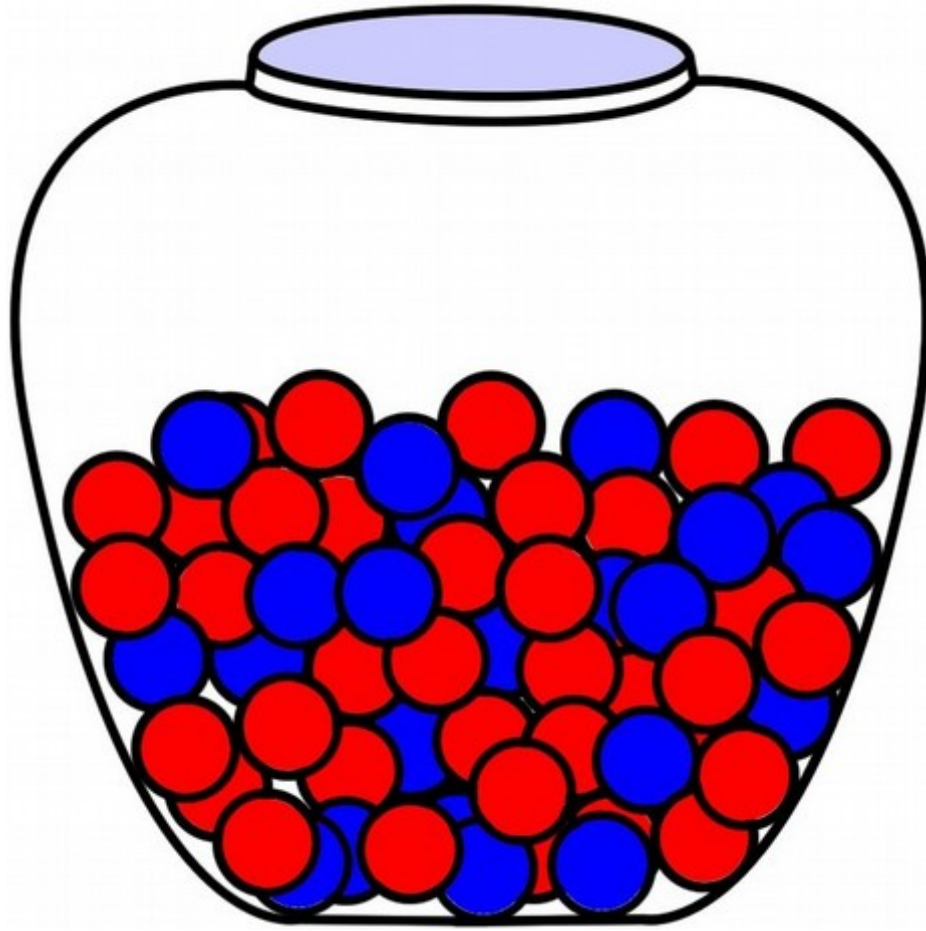
Muestra



- ν = fracción de canicas rojas

¿Qué nos dice ν de μ ?

Probabilidad al rescate



Respuesta 1

- Tomando cualquier N de canicas no se sabe nada de las que están en el frasco
- Se podría tomar una fracción grande de rojas y la realidad es que es al revés

Possible pero no probable

Probabilidad al rescate

Ejercicio

- Si $\mu = 0.9$, ¿cuál es la probabilidad de que una muestra de 10 tengan $v \leq 0.1$?

Probabilidad al rescate

Desigualdad de Hoeffding

$$\mathbb{P}[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}, \quad \forall \epsilon > 0$$

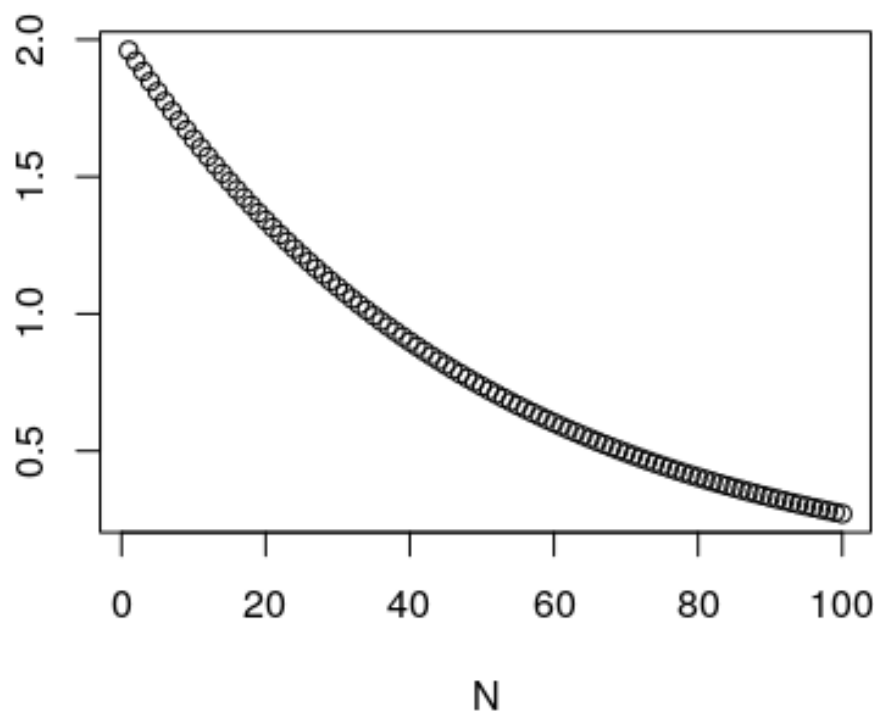
- Si el tamaño de la muestra N crece, se hace exponencialmente improbable que ν se desvíe de μ por más de la tolerancia ϵ .

Probabilidad al rescate

Desigualdad de Hoeffding

$$\mathbb{P}[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}, \quad \forall \epsilon > 0$$

$\epsilon=0.1$



Probabilidad al rescate

Desigualdad de Hoeffding

$$\mathbb{P}[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}, \quad \forall \epsilon > 0$$

- ν es aleatorio
- μ no, aunque es desconocido

Importante

- La desigualdad sirve para inferir μ usando ν aunque μ afecta ν y no viceversa
- Ya que el efecto es que ν tiende a estar cerca a μ , inferimos que μ tiende a estar cerca de ν

Probabilidad al rescate

Desigualdad de Hoeffding

$$\mathbb{P}[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}, \quad \forall \epsilon > 0$$

Importante

- Aunque $\mathbb{P}[|\nu - \mu| > \epsilon]$ depende de μ , se puede acotar la probabilidad mediante $2e^{-2\epsilon^2 N}$ que no depende de μ
- Sólo el tamaño N de la muestra afecta al límite, no el tamaño del “frasco de canicas”

Probabilidad al rescate

Ejercicio

- Si $\mu = 0.9$, use la desigualdad de Hoeffding para acotar la probabilidad de que una muestra de 10 canicas tengan $v \leq 0.1$ y compare la respuesta con la respuesta del ejercicio anterior

Probabilidad al rescate

Relación canicas – problema del aprendizaje

- Tomando una hipótesis h perteneciente a \mathcal{H} y comparándola a f en cada punto \mathbf{x} perteneciente a \mathcal{X}
- Si $h(\mathbf{x}) = f(\mathbf{x})$, azul
- Si $h(\mathbf{x}) \neq f(\mathbf{x})$, roja
- El color de cada punto es desconocido porque no conocemos f (se habla de comparar todos los \mathbf{x} de \mathcal{X})

Probabilidad al rescate

- Ahora tomando un x aleatorio de acuerdo a alguna distribución de probabilidad P en el espacio de entrada (\mathcal{X}), se sabe que x será roja con probabilidad μ y azul con probabilidad $1-\mu$
- Ahora, sin importar el valor de μ , el espacio de entrada se comporta ahora como el modelo del frasco y las canicas

Probabilidad al rescate

- Los ejemplos de entrada serían como una muestra del frasco
- Si la entrada x_1, \dots, x_N en \mathcal{D} son tomados independientemente de acuerdo a P se tendrá una muestra de puntos rojos ($h(x_n) \neq f(x_n)$) y azules ($h(x_n) = f(x_n)$)
- Cada punto será rojo con probabilidad μ y azul con probabilidad $1-\mu$
- Ahora el color de cada punto será conocido porque $h(x_n)$ y $f(x_n)$ son conocidos para $n=1, \dots, N$

Probabilidad al rescate

- El problema general de aprendizaje se puede reducir entonces al problema de las canicas

Suposición

- La entrada en \mathcal{D} es tomada independientemente de acuerdo a alguna distribución P en \mathcal{X}
- Cualquier P se puede traducir a un μ en el modelo de las canicas equivalente
- Ya que μ puede ser desconocido, P puede ser desconocido también

UNKNOWN TARGET FUNCTION

$$f: \mathcal{X} \rightarrow \mathcal{Y}$$

**PROBABILITY
DISTRIBUTION**

P on \mathcal{X}

TRAINING EXAMPLES

$$(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)$$

$$\mathbf{x}_1, \dots, \mathbf{x}_N$$

**LEARNING
ALGORITHM**

\mathcal{A}

**FINAL
HYPOTHESIS**

$$g \approx f$$

HYPOTHESIS SET

\mathcal{H}

Probabilidad al rescate

- Con esta analogía, se puede utilizar la desigualdad de Hoeffding permitiendo hacer una predicción fuera de \mathcal{D}
- Usando v para predecir μ dice algo acerca de f , aunque no dice cuál es f
- Lo que μ dice es la tasa de error que h hace al aproximar f
- Si v es cercano a cero, se puede predecir que h aproximará a f bien sobre **TODO** el espacio de entrada

Probabilidad al rescate

El problema

- No se tiene control sobre v puesto que se basa en una hipótesis h particular
- En la vida real se busca en todo el hypothesis set \mathcal{H}
- Si se tiene una sola hipótesis no se está aprendiendo sino sólo verificando si la hipótesis es buena o mala

Probabilidad al rescate

Extendiendo el modelo para múltiples hipótesis

In-sample error

- Tasa de error en la muestra \mathcal{D}
- Corresponde a v en el modelo de las canicas

$$\begin{aligned} E_{\text{in}}(h) &= (\text{fraction of } \mathcal{D} \text{ where } f \text{ and } h \text{ disagree}) \\ &= \frac{1}{N} \sum_{n=1}^N \mathbb{I}[h(\mathbf{x}_n) \neq f(\mathbf{x}_n)], \end{aligned}$$

- $\mathbb{I}[\cdot] = 1$ si T, $\mathbb{I}[\cdot] = 0$ si F.
- Dependencia explícita de E_{in} en un h particular

Probabilidad al rescate

Extendiendo el modelo para múltiples hipótesis

Out-of-sample error

- Tasa de error en todo el frasco
- Corresponde a μ en el modelo de las canicas

$$E_{\text{out}}(h) = \mathbb{P} [h(\mathbf{x}) \neq f(\mathbf{x})]$$

- Probabilidad basada en la distribución P sobre \mathcal{X} que es usada para sacar la muestra de data points \mathbf{x}

Probabilidad al rescate

Extendiendo el modelo para múltiples hipótesis

$$\mathbb{P}[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}, \quad \forall \epsilon > 0$$

$$\mathbb{P}[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2e^{-2\epsilon^2 N}, \quad \forall \epsilon > 0$$

- E_{in} es una **variable** aleatoria que depende de la muestra
- E_{out} es un **valor** desconocido pero no aleatorio

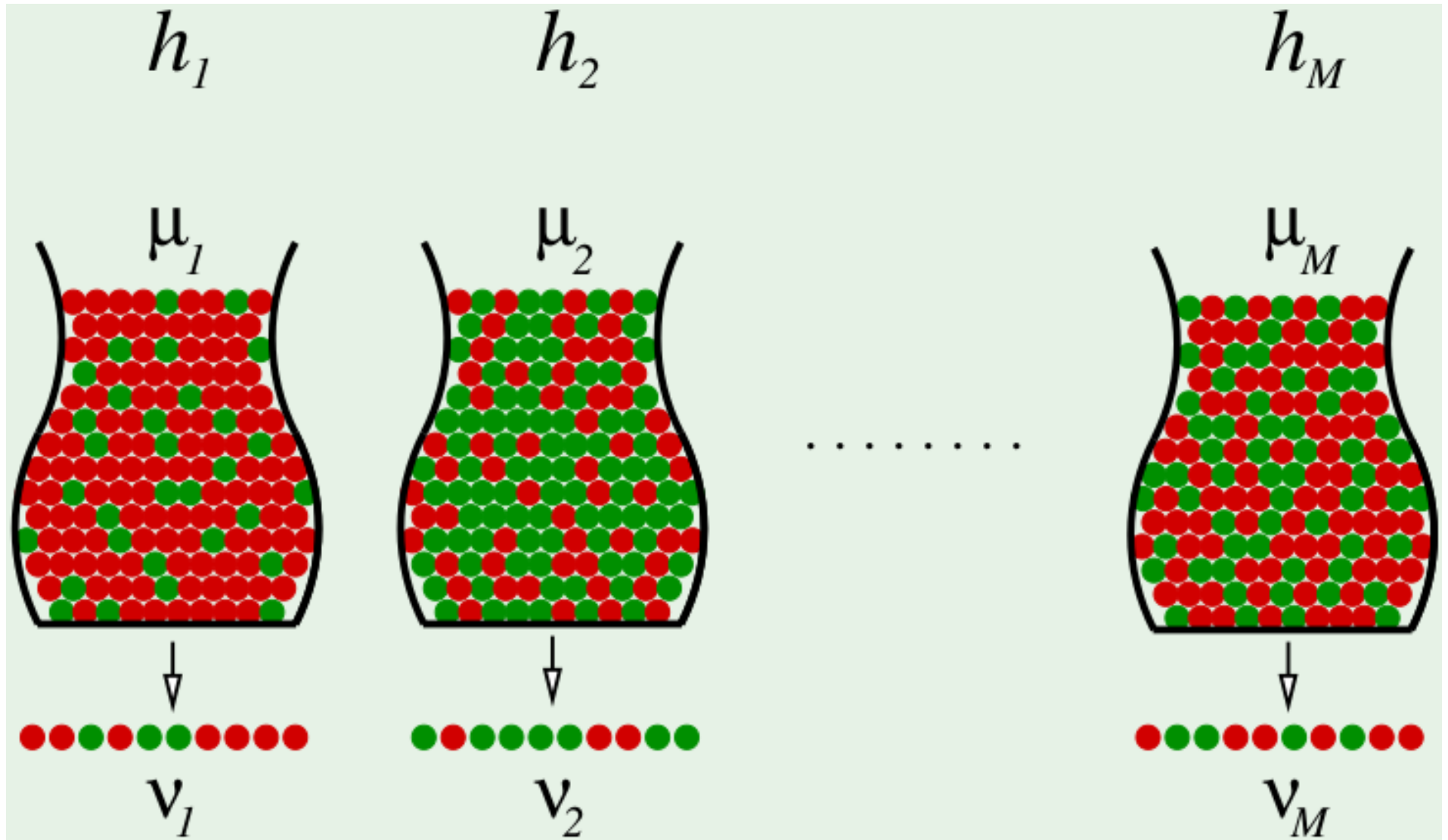
Probabilidad al rescate

Extendiendo el modelo para múltiples hipótesis

- Considere ahora el hypothesis set \mathcal{H} completo
- Asuma por el momento un \mathcal{H} un conjunto finito de M hipótesis

$$\mathcal{H} = \{h_1, h_2, \dots, h_M\}$$

Probabilidad al rescate



Probabilidad al rescate

- Aunque la desigualdad de Hoeffding se aplica aún para cada frasco individualmente, la situación es más complicada al considerar los frascos simultáneamente

El porqué

- En

$$\mathbb{P}[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2e^{-2\epsilon^2 N}, \quad \forall \epsilon > 0$$

- La hipótesis h es fijada antes de generar los datos de ejemplo y la probabilidad es con respecto al data set aleatorio \mathcal{D}

Probabilidad al rescate

El porqué

- En

$$\mathbb{P}[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2e^{-2\epsilon^2 N}, \quad \forall \epsilon > 0$$

- La hipótesis h es fijada antes de generar los datos de ejemplo y la probabilidad es con respecto al data set aleatorio \mathcal{D}
- Para la aplicación de la desigualdad de Hoeffding la suposición es obligatoria “ h es fijada antes de generar el data set \mathcal{D} ”

Probabilidad al rescate

- Con múltiples hipótesis en \mathcal{H} , el algoritmo de aprendizaje toma la hipótesis final g basada en \mathcal{D}
- **I.e. después de generar el data set de ejemplo = ¡no se cumple la suposición!**
- Y no se desea decir
“ $\mathbb{P}[|E_{in}(h) - E_{out}(h)| > \epsilon]$ es pequeña”
- Para cualquier h_m fijo perteneciente a \mathcal{H} particular

Probabilidad al rescate

- Lo que se desea decir es

“ $\mathbb{P}[|E_{in}(h) - E_{out}(h)| > \epsilon]$ es pequeña”

- para la hipótesis final g
- La hipótesis g no es fijada antes de generar los datos porque justamente qué hipótesis es seleccionada para ser g depende de los datos
- No se cumple la suposición y no se puede directamente poner g en lugar de h en la desigualdad de Hoeffding

Probabilidad al rescate

Exercise 1.10

Here is an experiment that illustrates the difference between a single bin and multiple bins. Run a computer simulation for flipping 1,000 fair coins. Flip each coin independently 10 times. Let's focus on 3 coins as follows: c_1 is the first coin flipped; c_{rand} is a coin you choose at random; c_{min} is the coin that had the minimum frequency of heads (pick the earlier one in case of a tie). Let ν_1 , ν_{rand} and ν_{min} be the fraction of heads you obtain for the respective three coins.

- (a) What is μ for the three coins selected?
- (b) Repeat this entire experiment a large number of times (e.g., 100,000 runs of the entire experiment) to get several instances of ν_1 , ν_{rand} and ν_{min} and plot the histograms of the distributions of ν_1 , ν_{rand} and ν_{min} . Notice that which coins end up being c_{rand} and c_{min} may differ from one run to another.
- (c) Using (b), plot estimates for $\mathbb{P}[|\nu - \mu| > \epsilon]$ as a function of ϵ , together with the Hoeffding bound $2e^{-2\epsilon^2 N}$ (on the same graph).
- (d) Which coins obey the Hoeffding bound, and which ones do not? Explain why.
- (e) Relate part (d) to the multiple bins in Figure 1.10.

Probabilidad al rescate

- Lo que se quiere es delimitar

$$\mathbb{P}[|E_{in}(h) - E_{out}(h)| > \epsilon]$$

- de forma que no se dependa de qué g el algoritmo escoge


La solución

- Ya que g debe ser uno de los h_m 's sin importar el algoritmo y la muestra, siempre se cumple que

Probabilidad al rescate

$$\begin{aligned} \text{"}|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\text{"} &\implies \text{"} |E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| > \epsilon \\ &\text{or } |E_{\text{in}}(h_2) - E_{\text{out}}(h_2)| > \epsilon \\ &\dots \\ &\text{or } |E_{\text{in}}(h_M) - E_{\text{out}}(h_M)| > \epsilon \text{"} \end{aligned}$$

Propiedad deseada:
las hipótesis h_m 's son fijas



Probabilidad al rescate

Regla de probabilidad

if $\mathcal{B}_1 \implies \mathcal{B}_2$, then $\mathbb{P}[\mathcal{B}_1] \leq \mathbb{P}[\mathcal{B}_2]$

Union bound

$$\mathbb{P}[\mathcal{B}_1 \text{ or } \mathcal{B}_2 \text{ or } \cdots \text{ or } \mathcal{B}_M] \leq \mathbb{P}[\mathcal{B}_1] + \mathbb{P}[\mathcal{B}_2] + \cdots + \mathbb{P}[\mathcal{B}_M]$$

- Usando las dos reglas se tiene que

Probabilidad al rescate

$$\begin{aligned}\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] &\leq \mathbb{P}[|E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| > \epsilon \\ &\quad \text{or } |E_{\text{in}}(h_2) - E_{\text{out}}(h_2)| > \epsilon \\ &\quad \dots \\ &\quad \text{or } |E_{\text{in}}(h_M) - E_{\text{out}}(h_M)| > \epsilon] \\ &\leq \sum_{m=1}^M \mathbb{P}[|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon].\end{aligned}$$

$$\mathbb{P}[|E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}, \quad \forall \epsilon > 0$$

Probabilidad al rescate

$$\mathbb{P}[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}, \quad \forall \epsilon > 0$$

- Lo que se trata es de simultáneamente aproximar todos los $E_{out}(h_m)$'s mediante los correspondientes $E_{in}(h_m)$'s.
- Esto permite al algoritmo de aprendizaje escoger cualquier hipótesis de E_{in} y esperar que el correspondiente E_{out} uniformemente siga el juego sin tener en cuenta qué hipótesis es escogida

Factibilidad del aprendizaje

Factibilidad del aprendizaje

Argumentos contrarios

- No se puede aprender nada fuera de \mathcal{D}
- Sí se puede aprender algo fuera de \mathcal{D}

Factibilidad del aprendizaje

Conciliar los dos argumentos

- ¿ \mathcal{D} dice algo fuera de \mathcal{D} que no se sabía antes?
- Determinísticamente
 - No. Algo cierto acerca de f
- Probabilísticamente
 - Probablemente algo acerca de f

Suposición

- \mathcal{D} es generado independientemente

Factibilidad del aprendizaje

Exercise 1.11

We are given a data set \mathcal{D} of 25 training examples from an unknown target function $f: \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \{-1, +1\}$. To learn f , we use a simple hypothesis set $\mathcal{H} = \{h_1, h_2\}$ where h_1 is the constant $+1$ function and h_2 is the constant -1 .

We consider two learning algorithms, S (smart) and C (crazy). S chooses the hypothesis that agrees the most with \mathcal{D} and C chooses the other hypothesis deliberately. Let us see how these algorithms perform out of sample from the deterministic and probabilistic points of view. Assume in the probabilistic view that there is a probability distribution on \mathcal{X} , and let $\mathbb{P}[f(\mathbf{x}) = +1] = p$.

- (a) Can S produce a hypothesis that is *guaranteed* to perform better than random on any point outside \mathcal{D} ?
- (b) Assume for the rest of the exercise that all the examples in \mathcal{D} have $y_n = +1$. Is it *possible* that the hypothesis that C produces turns out to be better than the hypothesis that S produces?
- (c) If $p = 0.9$, what is the probability that S will produce a better hypothesis than C ?
- (d) Is there any value of p for which it is more likely than not that C will produce a better hypothesis than S ?

Factibilidad del aprendizaje

- ¿Qué significa factibilidad del aprendizaje?
 - 1) ¿Podemos estar seguros que $E_{\text{out}}(g)$ es suficientemente cerca a $E_{\text{in}}(g)$?
 - 2) ¿Se puede hacer $E_{\text{in}}(g)$ suficientemente pequeño?
- No siempre se puede asegurar $E_{\text{in}}(g)$

Factibilidad del aprendizaje

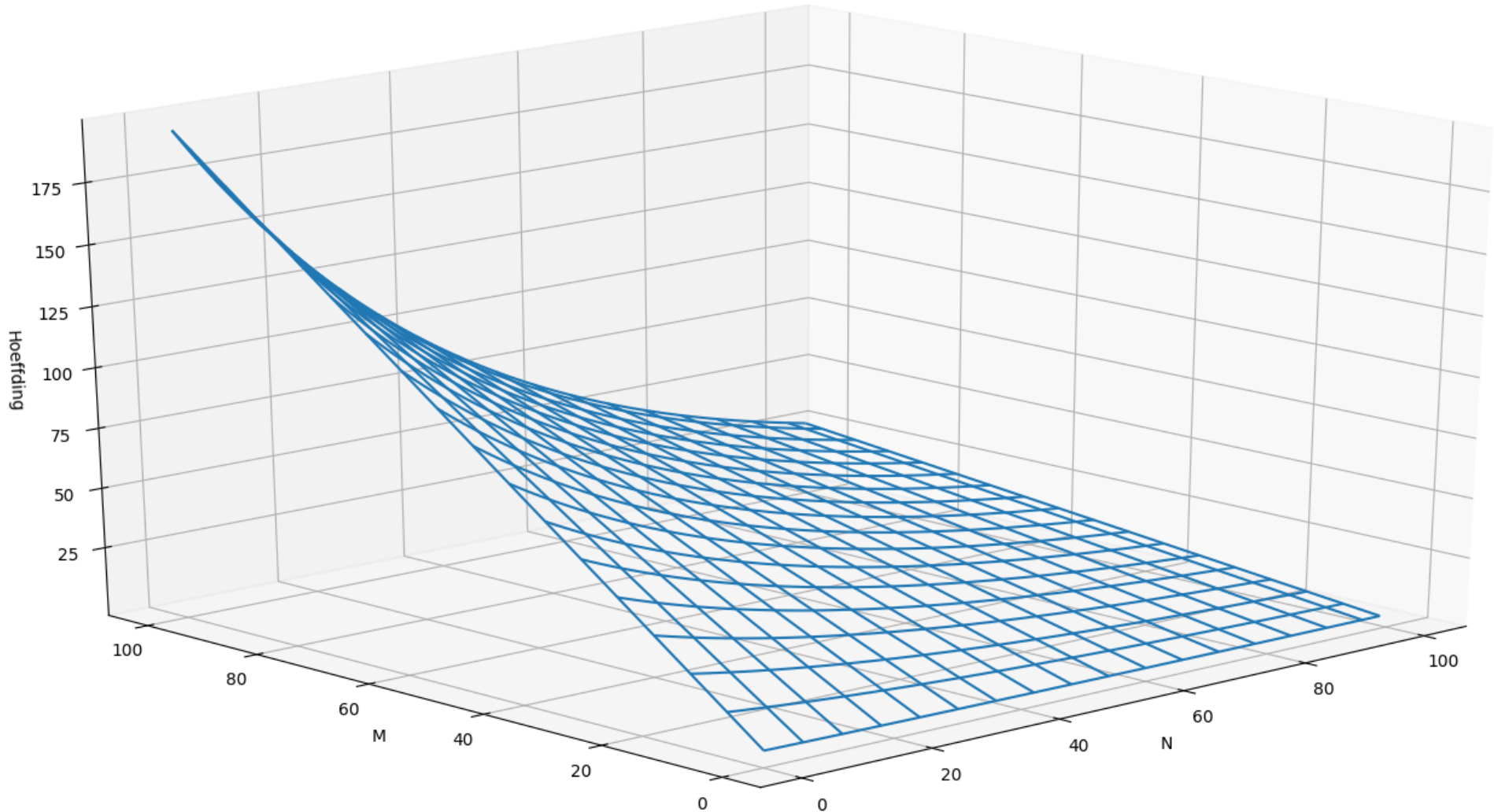
Papel de cada componente en el problema del aprendizaje

La complejidad de H

- Si M crece, hay más riesgo de que $E_{\text{in}}(g)$ sea un estimador pobre de $E_{\text{out}}(g)$
- M podría ser una medida de la complejidad de \mathcal{H}
- Si se desea que $E_{\text{in}}(g) \approx E_{\text{out}}(g)$ se necesita tomar mucho en cuenta M
- Si se desea un $E_{\text{in}}(g)$ pequeño es bueno tener un M grande para tener más posibilidad de encontrar un g que se acople bien a los datos (\mathcal{H} más complejo)
- Tradeoff en la complejidad de \mathcal{H} , teoría del aprendizaje

Factibilidad del aprendizaje

$\varepsilon = 0.1$



Factibilidad del aprendizaje

Papel de cada componente en el problema del aprendizaje

La complejidad de f

- Intuitivamente un f más complejo debería ser más difícil de aprender que un f menos complejo

$$\mathbb{P}[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}, \quad \forall \epsilon > 0$$

- La complejidad de f no afecta qué tan bien $E_{in}(g)$ aproxima $E_{out}(g)$

Factibilidad del aprendizaje

Papel de cada componente en el problema del aprendizaje

La complejidad de f

$$\mathbb{P}[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}, \quad \forall \epsilon > 0$$

- La complejidad de f no afecta qué tan bien $E_{in}(g)$ aproxima $E_{out}(g)$
- ¿Significa que es igual de fácil/difícil aprender f 's simples que f 's complejas?

Factibilidad del aprendizaje

Papel de cada componente en el problema del aprendizaje

La complejidad de f

$$\mathbb{P}[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}, \quad \forall \epsilon > 0$$

- Si f es compleja es más complejo ajustar h a \mathcal{D}
- Peor $E_{in}(g)$ cuando f es complejo
- Hacer el \mathcal{H} más complejo para ajustar mejor los datos (E_{in} bajo) pero E_{out} se alejará más de E_{in}

Conceptos y términos importantes

Conceptos y términos importantes

- Desigualdad de Hoeffding y su supuesto
- \mathcal{D} es generado independientemente
- Trade-off complejidad de \mathcal{H} , $E_{\text{out}}/E_{\text{in}}$

Referencias

- Abu-Mostafa, Y.S., Magdon-Ismael, M., Lin, H.-T., 2012. Learning from data: a short course. AMLbook.com, USA.
- Raschka, S., 2016. Python machine learning, Community experience distilled. Packt Publishing, Birmingham Mumbai.

Preguntas

