

Monografía sobre Aprendizaje Automático

Nicolás Caro

Departamento de Ingeniería Matemática, Universidad de Chile, Santiago.

`ncaro@dim.uchile.cl`

August 9, 2018

Contents

Preface	vii
Chapter 1. Modelos gráficos probabilísticos	1
Introducción	1
1.1. Modelos gráficos dirigidos	1
1.1.1. Naive Bayes	3
1.1.2. Regresión polinomial	3
1.1.3. Modelos gráficos dirigidos gaussianos	4
1.2. Independencia condicional en modelos gráficos dirigidos	5
1.2.1. d-separación	6
1.2.2. Markov blankets	7
1.3. Modelos gráficos no dirigidos	8
1.4. Inferencia exacta en modelos gráficos	8
Chapter 2. Métodos de inferencia aproximada	9
Inferencia Monte Carlo	9
Inferencia Markov chain Monte Carlo	9
Inferencia variacional	9
Chapter 3. Aprendizaje con Kernels	11
3.1. Introducción	11
3.2. Terminología y propiedades	11
Chapter 4. Procesos Gaussianos	15
Introducción	15
4.1. Procesos Gaussianos	15
Chapter 5. Redes Neuronales	17
Bibliography	19
Index	21

Preface

This document is a sample prepared to illustrate the use of the American Mathematical Society's L^AT_EX document class `amsbook` and publication-specific variants of that class.

This is an example of an unnumbered chapter which can be used for a Preface or Foreword.

The purpose of this paper is to establish a relationship between an infinite-dimensional Grassmannian and arbitrary algebraic vector bundles of any rank defined over an arbitrary complete irreducible algebraic curve, which generalizes the known connection between the Grassmannian and line bundles on algebraic curves.

Author Name

Modelos gráficos probabilísticos

El *machine learning* fue un objeto de lujo, pero para nosotros es un artículo de primera necesidad: no podemos vivir sin *machine learning*.

NicoCaro:
Poner algo profundo por el estilo. (?)

Introducción

El eje central de este capítulo se basa en la búsqueda de una representación compacta, para distribuciones de probabilidad conjunta de la forma $p(\mathbf{x}|\boldsymbol{\theta})$. Esto, con la intención de realizar inferencia sobre variables y aprendizaje de parámetros de manera eficiente.

NicoCaro:
Mejorar intro, añadir discusión sobre inferencia y aprendizaje.

1.1. Modelos gráficos dirigidos

Toda distribución de probabilidad conjunta $p(\mathbf{x}) = p(x_1, x_2, \dots, x_v)$ se puede representar de la forma:

$$(1.1) \quad p(\mathbf{x}) = p(x_1)p(x_2 | x_1)p(x_3 | x_1, x_2) \dots p(x_v | x_1, x_2, \dots, x_{v-1})$$

El problema con esta expresión es la dificultad computacional subyacente al cálculo de distribuciones condicionales de la forma $p(x_t | x_1, \dots, x_{t-1})$ cuando el número de variables incidentes t aumenta.

NicoCaro:
discusión sobre computabilidad. (?) (ref: Murphy, pg. 307)

No obstante, la representación (1.1) reduce su complejidad en presencia de **independencia condicional**.

En efecto, si se asume $x_{t+1} \perp x_1, \dots, x_{t-1} | x_t$. Es decir, las observaciones futuras x_{t+1} son independientes del pasado x_1, \dots, x_{t-1} , dado el estado presente x_t . La probabilidad conjunta se reduce entonces a:

$$(1.2) \quad p(\mathbf{x}) = p(x_1) \prod_{t=2}^v p(x_t | x_1, \dots, x_{t-1}) = p(x_1) \prod_{t=2}^v p(x_t | x_{t-1})$$

NicoCaro:
propiedades básicas del cálculo de probabilidades (CI por ej.) al apéndice.

De lo cual se obtiene una expresión más simple.

Modelar la independencia condicional entre las variables permite entonces reducir la complejidad de representación para la distribución conjunta. En particular, la elección tomada en (1.2) se conoce como **propiedad de Markov** de primer orden. En un contexto general, las relaciones de independencia condicional entre variables aleatorias de dimensión arbitraria, se modelan utilizando *diagramas de independencia* o **modelos gráficos**. Estos se valen de un grafo $G = (\mathcal{V}, \mathcal{E})$ ¹ para representar mediante nodos $v = 1, \dots, \mathcal{V}$ las variables aleatorias del modelo, mientras que la presencia o ausencia de aristas entre estos nodos, permite modelar las relaciones de dependencia condicional subyacentes.

¹Conjunto consistente de $\mathcal{V} = \{1 \dots, V\}$ vértices (o nodos) y $\mathcal{E} = \{(s, t) : s, t \in \mathcal{V}\}$ aristas.

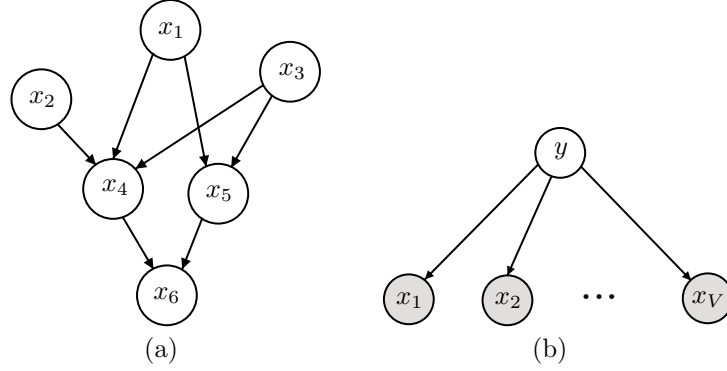


FIGURE 1. (a) Ejemplo de modelo gráfico dirigido. (b) Relaciones de dependencia condicional en el clasificador naive Bayes como un modelo gráfico dirigido, las variables aleatorias observadas se denotan por nodos grises.

Una *red bayesiana* o **modelo gráfico dirigido** es un modelo gráfico probabilístico, cuyo grafo subyacente es un **grafo dirigido acíclico** (DAG por sus siglas en inglés). Todo DAG posee un *ordenamiento topológico*, es decir, los nodos de cualquier DAG pueden ser numerados de manera tal, que todo nodo padre posea una numeración inferior a sus nodos hijos. Esta característica permite enriquecer la formulación de la propiedad de Markov (1.2), usando la estructura grafica como componente adicional. De esta forma, se puede formular la **propiedad ordenada de Markov** en modelos gráficos dirigidos:

$$(1.3) \quad x_s \perp \mathbf{x}_{pred(s) \setminus pa(s)} \mid \mathbf{x}_{pa(s)}$$

Es decir, un nodo x_s es independiente de aquellos predecesores, menores en orden topológico, a sus padres $\mathbf{x}_{pred(s) \setminus pa(s)}$, dados sus nodos padres $\mathbf{x}_{pa(s)}$. De manera equivalente, un nodo x_s solo depende de sus padres inmediatos $x_{pa(s)}$ y no de todos sus predecesores.

De esta forma, la probabilidad conjunta de un modelo gráfico dirigido, que cumple la propiedad ordenada de Markov, se puede descomponer de la forma:

$$(1.4) \quad p(\mathbf{x}) = \prod_{t=1}^V p(x_t | \mathbf{x}_{pa(t)})$$

EXAMPLE 1.1 (Modelo grafico asociado a $p(\mathbf{x})$). Si se estudia un modelo probabilístico, donde la probabilidad conjunta de las variables estudiadas $p(\mathbf{x})$ esta dada por:

$$(1.5) \quad p(\mathbf{x}) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4, x_5)$$

Entonces, un grafo dirigido asociado a tal factorización es el de la figura 1(a). Para construir dicho grafo, se consideran las relaciones de independencia condicional en la factorización (1.5), para luego establecer aristas $s \rightarrow t$ si la probabilidad condicional del nodo x_s depende de x_t . En este caso, no hay aristas incidentes hacia x_1 , x_2 ni x_3 . Por otra parte, se deben crear aristas desde x_1, x_2 y x_3 hacia x_4 , desde x_1 y x_3 hacia x_5 y desde x_4, x_5 hacia x_6 .

NicoCaro:
lema de or-
denamiento
topológico
para DAG's en
apéndice.

En general, es posible reconstruir la probabilidad conjunta subyacente a un modelo gráfico probabilístico conociendo el grafo y haciendo el proceso inverso al descrito anteriormente.

Con el fin de explorar las posibilidades de este tipo de modelos e introducir conceptos referentes a la notación de estos, se pasan a estudiar los siguientes ejemplos:

1.1.1. Naive Bayes. Dado un problema de clasificación de vectores $\mathbf{x} = (x_1, \dots, x_V)$ en C clases. Es posible modelar las variables de decisión x_t como condicionalmente independientes dada la categoría de clasificación:

$$(1.6) \quad x_i \perp x_j \mid y = c, \quad i \neq j$$

Si se usa este enfoque, se obtiene que la densidad condicional de clases toma la forma:

$$(1.7) \quad p(\mathbf{x} \mid y = c) = \prod_{t=1}^V p(x_t \mid y = c)$$

Al parametrizar las distribuciones de densidad condicional, es posible obtener un modelo de clasificación conocido como **clasificador naive Bayes**. La estructura de las relaciones de independencia inducidas por (1.6) se pueden expresar según (1.7) y el modelo gráfico dirigido de la figura 1(b).

1.1.2. Regresión polinomial. las variables aleatorias son el vector de coeficientes polinomiales \mathbf{w} y los datos observados $\mathbf{y} = (y_1, \dots, y_N)^T$. Adicionalmente, se parametriza el ruido del modelo a través de σ_ε^2 y la varianza de la distribución a priori ² de \mathbf{w} por σ_w^2 . Finalmente, los datos de entrada se denotan por $\mathbf{x} = (x_1, \dots, x_N)^T$.

NicoCaro:
añadir intro significativa

La probabilidad conjunta de este modelo es el producto de la probabilidad a priori $p(\mathbf{w})$ con las distribuciones condicionales $p(y_i \mid \mathbf{w})$ para $i = 1, \dots, N$:

$$(1.8) \quad p(\mathbf{y}, \mathbf{w}) = p(\mathbf{w}) \prod_{i=1}^N p(y_i \mid \mathbf{w})$$

El grafo de tal factorización es similar al del clasificador naive Bayes 1(b). Para representarlo de manera compacta, se usa la notación de de placas o *plates*, en la figura 2(a) se muestra el grafo de (1.8) usando esta convención. Aquí N es la cantidad de nodos del modelo, de los cuales se muestra el representante y_i .

Si por otra parte, si se quiere estudiar la interacción de los parámetros en el modelo, es posible explicitarlos en la probabilidad conjunta para luego agregarlos al grafo:

$$(1.9) \quad p(\mathbf{y}, \mathbf{w} \mid \mathbf{x}, \sigma_\varepsilon^2, \sigma_w^2) = p(\mathbf{w} \mid \sigma_w^2) \prod_{i=1}^N p(y_i \mid \mathbf{w}, x_i, \sigma_\varepsilon^2)$$

La figura 2(b) muestra el grafo correspondiente a (1.9). Por convención, las variables deterministas se incluyen en el grafo como círculos pequeños, mientras que las variables aleatorias observadas se muestran como nodos grises, los nodos incoloros representan variables latentes o no observadas, finalmente las aristas, al

²Considerándose una distribución a priori, gaussiana y esférica sobre \mathbf{w} .

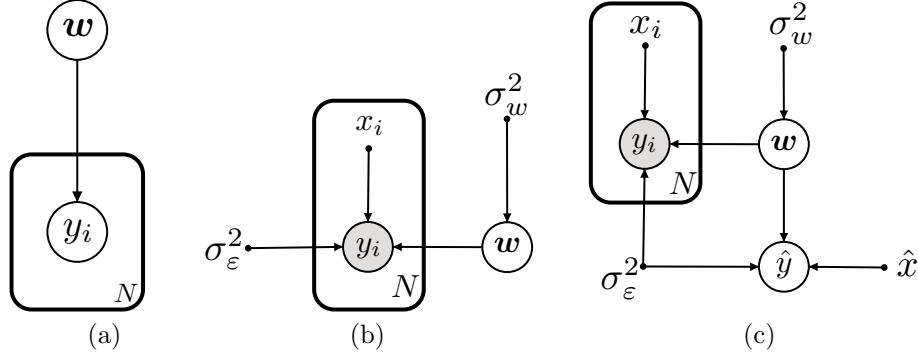


FIGURE 2. Modelo grafico dirigido para regresión polinomial usando notación de placas (o *plates*). En (a) se muestra el grafo correspondiente a (1.8). En (b) se añaden los parámetros deterministas y las variables aleatorias observadas. En (c) se añaden datos de entrada y predicciones.

NicoCaro:
ver si es necesario
cambiar el
formato de los
3 grafos juntos.
(muy pegados ?)

igual que en los ejemplos anteriores, representan la dependencia condicional en la factorización de la probabilidad conjunta.

Para realizar predicciones en datos nuevos \hat{x} , se desea encontrar la distribución de probabilidad para \hat{y} condicionada a la información que ya se posee. Esta corresponde a:

$$(1.10) \quad p(\hat{y}, \mathbf{y}, \mathbf{w} | \hat{x}, \mathbf{x}, \sigma_w^2, \sigma_\varepsilon^2) = \left[\prod_{i=1}^N p(y_i | x_i, \mathbf{w}, \sigma_\varepsilon^2) \right] p(\mathbf{w} | \sigma_w^2) p(\hat{y} | \hat{x}, \mathbf{w}, \sigma_\varepsilon^2)$$

Finalmente, se deduce la distribución predictiva para \hat{y} :

$$(1.11) \quad p(\hat{y} | \hat{x}, \mathbf{y}, \mathbf{x}, \sigma_w^2, \sigma_\varepsilon^2) \propto \int p(\hat{y}, \mathbf{y}, \mathbf{w} | \hat{x}, \mathbf{x}, \sigma_w^2, \sigma_\varepsilon^2) d\mathbf{w}$$

El modelo gráfico dirigido que encapsula estas últimas ecuaciones se aprecia en 2(c).

1.1.3. Modelos gráficos dirigidos gaussianos. Sea \mathcal{M} un modelo grafico dirigido, en el cual todas las variables son reales y sus distribuciones de probabilidad condicional son lineal-gaussianas:

$$(1.12) \quad p(x_t | \mathbf{x}_{pa(t)}) = \mathcal{N}(x_t | \mu_t + \mathbf{w}_t^T \mathbf{x}_{pa(t)}, \sigma_t^2)$$

La estructura de \mathcal{M} permite modelar la probabilidad conjunta de las variables del modelo en la forma:

$$(1.13) \quad p(\mathbf{x} | \mathcal{M}) = \prod_{t=1}^V p(x_t | \mathbf{x}_{pa(t)}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Lo cual se conoce como **red bayesiana gaussiana**. Para este tipo de modelos, es posible inferir $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$. En efecto, según (1.13):

$$(1.14) \quad \log p(\mathbf{x} | \mathcal{M}) = - \sum_{t=1}^V \frac{1}{2\sigma_t^2} \left(x_t - \sum_{s \in pa(t)} w_{ts} x_s - \mu_t \right)^2 + K$$

Donde K representa una constante independiente de \mathbf{x} . Al ser la log-probabilidad conjunta, cuadrática en las componentes de \mathbf{x} , se obtiene que efectivamente la probabilidad conjunta es normal multivariada para \mathbf{x} en (1.13). Para estimar la media, se observa en primera instancia:

$$(1.15) \quad x_t = \sum_{s \in pa(t)} w_{ts} \mathbb{E}[x_s] + \mu_t + \sigma_t \varepsilon_t$$

Donde $\varepsilon_t \sim \mathcal{N}(0, 1)$ y $\mathbb{E}[\varepsilon_t, \varepsilon_s] = 0$, para $s \neq t$. De esto se deduce:

$$(1.16) \quad \mathbb{E}[x_t] = \sum_{s \in pa(t)} w_{ts} \mathbb{E}[x_s] + \mu_t$$

Es posible entonces, encontrar las componentes de $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}] = (\mathbb{E}[x_1], \dots, \mathbb{E}[x_V])^T$ utilizando la estructura gráfica dirigida de \mathcal{M} (y por tanto su ordenamiento topológico). Para ello, se comienza calculando $\mathbb{E}[x_1]$ para luego continuar de manera recursiva según la numeración de los nodos.

Similarmente, es posible calcular el elemento Σ_{st} de la matriz de covarianza, observando:

$$(1.17) \quad \begin{aligned} \text{cov}(x_s, x_t) &= \mathbb{E}[(x_s - \mathbb{E}[x_s])(x_t - \mathbb{E}[x_t])] \\ &= \left[(x_s - \mathbb{E}[x_s]) \left\{ \sum_{k \in pa(x_t)} w_{tk} (x_k - \mathbb{E}[x_k]) + \sigma_t \varepsilon_t \right\} \right] \\ &= \sum_{k \in pa(x_t)} w_{tk} \text{cov}[x_s, x_k] + \sigma_t^2 \mathbf{I}_{st} \end{aligned}$$

De donde al igual que en (1.16), se calculan los elementos de Σ recursivamente.

Finalmente, se puede extender el modelo inducido por (1.12) a uno donde los nodos del modelo gráfico representen variables aleatorias gaussianas multivariantes. Para esto, se reescribe la distribución de probabilidad condicional para el nodo x_t en la forma:

$$(1.18) \quad p(\mathbf{x}_t \mid pa(\mathbf{x}_t)) = \mathcal{N} \left(\mathbf{x}_t \mid \sum_{s \in pa(\mathbf{x}_t)} \mathbf{W}_{ts} \mathbf{x}_s + \boldsymbol{\mu}_t, \Sigma_t \right)$$

Donde \mathbf{W}_{ts} es una matriz de pesos entre los vectores de cada nodo.

1.2. Independencia condicional en modelos gráficos dirigidos

Como se mencionó anteriormente, los modelos gráficos encapsulan las relaciones de independencia condicional entre las variables aleatorias del fenómeno que se modela. En esta sección se estudian las propiedades de los modelos gráficos dirigidos en cuanto sus propiedades

En un grafo G , se escribe $x_i \perp_G x_j \mid x_k$ si el nodo x_i es independiente de x_j dado x_k . Se denota por $I(G)$ al conjunto de todas las relaciones de independencia condicional codificadas en el grafo G .

DEFINITION 1.2 (Diagrama de independencia). Sea $p(\cdot)$ una distribución de probabilidad. Se dice que un grafo G es un diagrama de independencia o *I-map* para p si y solo si $I(G) \subseteq I(p)$. Donde $I(p)$ es el conjunto de todas las relaciones de independencia condicional ciertas para las variables de p .

NicoCaro:
terminar intro

De la definición anterior, se deduce que un grafo G es un diagrama de independencia para la distribución de probabilidad p , si este no contiene más relaciones de independencia condicional que las permitidas por p . De esta forma, toda distribución de probabilidad $p(\mathbf{x})$, donde $\mathbf{x} = (x_1, \dots, x_V)^T$, posee al menos un diagrama de independencia. En efecto, si se considera un grafo G con nodos $\mathcal{V} = \{x_1, \dots, x_V\}$ completamente conectados, entonces G es un diagrama de independencia para p pues no presenta aristas faltantes y por tanto se condiciona en todas las variables.

De la discusión anterior, tiene sentido hablar de un *diagrama de independencia minimal* G para p , es decir, un grafo G , tal que si G' es otro diagrama de independencia para p que cumple $G' \subseteq G$, entonces $G' = G$.

Finalmente, tal representación, permite extraer de su estructura gráfica, relaciones no triviales de independencia condicional, entre las variables de importancia. En el caso de un modelo gráfico dirigido, la noción de *separación dirigida* o *d-separación* facilita dicha tarea.

1.2.1. d-separación. Se dice que un *camino no dirigido* P está *separado de manera dirigida* o *d-separado* por un conjunto de nodos E , si y solo si, se cumple alguna de las siguiente condiciones:

- (1) P contiene una cadena, $s \rightarrow e \rightarrow t$, donde $e \in E$.
- (2) P contiene una estructura $s \leftarrow e \rightarrow t$, donde $e \in E$.
- (3) P contiene una estructura $s \rightarrow e \leftarrow t$, donde $e \notin E$ o e **no** es descendiente de algún elemento de E .

Se dice que un conjunto de nodos A está d-separado de un conjunto de nodos B , dado un conjunto de nodos E , si y solo si, todo camino no dirigido desde cada nodo de A a cada nodo de B está d-separado por E .

En un grafo acíclico dirigido G , se aprecia la siguiente propiedad:

Para comprender las propiedades anteriores, se analizan los siguientes ejemplos:

- Sea $x \rightarrow y \rightarrow z$ una cadena, tal grafo codifica la siguiente probabilidad conjunta:

$$p(x, y, z) = p(x)p(y|x)p(z|y)$$

Usando la propiedad (1), se puede deducir que $x \perp z|y$. Esto se comprueba pues:

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x)p(y|x)p(z|y)}{p(y)} = \frac{p(x)p(y)p(z|y)}{p(y)} = p(x|y)p(z|y)$$

- Sea la estructura $x \leftarrow y \rightarrow z$, según la propiedad (2), $x \perp z|y$. En efecto,

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(y)p(x|y)p(z|y)}{p(y)} = p(x|y)p(z|y)$$

- Sea finalmente la estructura $x \rightarrow y \leftarrow z$, en este caso $x \not\perp z|y$:

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x)p(y|x,z)p(z)}{p(y)} \neq p(x|y)p(z|y)$$

Es por tal motivo que en la propiedad (3), se requiere en este tipo de estructuras, que no existan nodos, ni descendientes de nodos de la familia condicionante E .

En el último caso, se puede comprobar que los nodos x e y son marginalmente independientes entre sí, es decir, $p(x, z) = p(x)p(z)$. Sin embargo, al condicionar ambos nodos por y , se vuelven dependientes, este efecto se denomina **paradoja de Berkson**. Finalmente, un modelo gráfico probabilístico que verifica la equivalencia (??) se dice que cumple la **propiedad global de Markov**.

1.2.2. Markov blankets. De la d-separación es posible concluir:

$$(1.19) \quad x_t \perp \mathbf{x}_{nd(t) \setminus pa(t)} | \mathbf{x}_{pa(t)}$$

Donde $nd(t)$ son los **no-descendientes** del nodo x_t ³. De esta forma, es posible concluir que en 1(a) $x_4 \perp x_5 | x_1, x_2, x_3$, pues en efecto, $nd(4) \setminus pa(4) = x_5$ y $pa(4) = x_1, x_2, x_3$. La ecuación (1.19) se conoce como **propiedad dirigida local de Markov**.

En especial, dado que $pred(t) \subseteq nd(t)$ se deriva la **propiedad ordenada de Markov**, ya presentada en (1.3). Sorprendentemente, estas tres propiedades son equivalentes.

Por otra parte, para cada nodo x_t es posible extraer el conjunto de nodos que lo separan del resto del grafo, es decir, se puede para cada nodo x_t , obtener el conjunto de todas las variables aleatorias que lo vuelven condicionalmente independiente a los demás nodos del modelo. El conjunto antes descrito se denomina **Markov blanket** y se denota por $mb(t)$ este conjunto de nodos corresponde a:

$$(1.20) \quad mb(t) := ch(t) \cup pa(t) \cup copa(t)$$

Donde $ch(t)$ son los nodos hijos de x_t , de manera análoga $pa(t)$ son nodos padres y $copa(t)$ sus copadres⁴. En la figura 1(a) se tiene por ejemplo $mb(5) = \{x_6, x_1, x_3, x_4\}$. Según la propiedad global de Markov, la presencia de los nodos copadres no parece ser necesaria en primera instancia (la dependencia condicional debería recaer únicamente en los nodos padres), sin embargo, al definir \mathbf{x}_{-t} como el conjunto de nodos distintos a x_t , es posible observar que la probabilidad conjunta adquiere la forma $p(\mathbf{x}) = p(x_t, \mathbf{x}_{-t})$. De donde, al marginalizar sobre el nodo x_t , se obtiene que $p(\mathbf{x}_{-t})$ contiene sólo a aquellos nodos del modelo en los que la variable x_t no aparece como argumento, ni como condicionante (dada la factorización de la probabilidad conjunta codificada en el grafo). Lo anterior implica que en $p(x_t | \mathbf{x}_{-t}) = p(\mathbf{x}) / p(\mathbf{x}_{-t})$ solo se podrán encontrar probabilidades condicionales donde x_t sea el argumento, lo que expresa con $p(x_t | \mathbf{x}_{pa(t)})$, o donde sea variable condicionante, es decir, sea padre o copadre de algún otro nodo. Se deduce:

$$(1.21) \quad p(x_t | \mathbf{x}_{-t}) \propto p(x_t | \mathbf{x}_{pa(t)}) \prod_{s \in ch(t)} p(x_s | \mathbf{x}_{pa(s)})$$

La expresión (1.21) se conoce como **condicional completa** del nodo x_t .

NicoCaro:
revisar de-
mostración Koller,
Friedman 2009

³ $nd(t) = \mathcal{V} \setminus \{t \cup desc(t)\}$, donde $desc(t)$ son los descendientes del nodo x_t , es decir, aquellos nodos que provienen de un camino dirigido con origen en x_t .

⁴ nodos que comparten hijos con x_t

1.3. Modelos gráficos no dirigidos

Los modelos gráficos dirigidos presentan una alternativa de modelación modular e interpretable. Sin embargo, su estructura los hace demasiado rígidos en aplicaciones donde las variables interactúan de manera simétrica (datos espaciales y relacionales por ejemplo).

Como alternativa, se encuentran los **modelos gráficos no dirigidos**, los cuales, como indica su nombre, no requieren el uso de aristas dirigidas y por tanto, encapsulan de manera natural la simetría que pierden los modelos dirigidos.

Para ilustrar la Introducción anterior, las figuras 3(a) y 3(b) muestran los markov blankets de un nodo para el caso dirigido y no dirigido en un arreglo 2d, común en aplicaciones referentes a imágenes y datos espaciales.

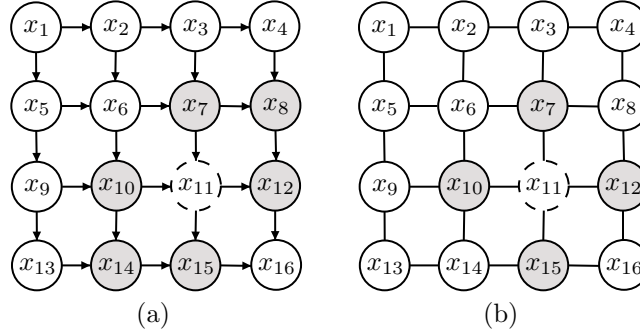


FIGURE 3. Modelo gráfico dirigido (izquierda) contrapuesto a uno no dirigido (derecha). En (a) el markov blanket de x_{11} añade a x_8 y x_{14} a sus vecinos inmediatos, mientras excluye a x_6 y x_{16} en aplicaciones basadas en imágenes, este comportamiento puede no ser el deseado. Por otra parte en (b) el nodo x_{11} esta separado del resto solo por sus vecinos inmediatos.

1.4. Inferencia exacta en modelos gráficos

CHAPTER 2

Métodos de inferencia aproximada

En este recinto se prohíbe dormir
Entrenar, validar, testear
Armonizar, huir, interceptar.

Inferencia Monte Carlo

En esta sección se estudian algoritmos basados en aproximación Monte Carlo, esto se basa en la obtención de muestras de una distribución de la forma $\mathbf{x}^s \sim p(\mathbf{x}|\mathcal{D})$ para calcular por ejemplo para la distribución posterior predictiva de cierto modelo $p(y|\mathcal{D})$ o cierta marginal posterior $p(x_1|\mathcal{D})$, esto se hace

Inferencia Markov chain Monte Carlo

Inferencia variacional

NicoCaro:
Poner algo profundo por el estilo(?)

FT:
estoy de acuerdo

Aprendizaje con Kernels

Alza del hiperparametro origina nueva alza del hiperparametro
 Alza de los errores
 Provoca instantáneamente la duplicación de los errores
 Alza de las métricas
 Origina alza de las métricas.

NicoCaro:
 Poner algo profundo por el estilo. (?)

3.1. Introducción

Un kernel es una función simétrica y definida positiva $k(\cdot, \cdot)$ que puede ser entendida como una medida de similitud entre los argumentos que opera. En el siguiente capítulo, se definen estos objetos matemáticos de manera formal, se investigan sus características y se derivan algunos métodos del aprendizaje de máquinas que toman ventaja sus propiedades.

3.2. Terminología y propiedades

El término **kernel** proviene del estudio de operadores integrales en el campo del análisis funcional. En tal contexto se les identifica como aquellas funciones k que determinan un operador T_k a través de:

$$(3.1) \quad (T_k f)(x) = \int_{\mathcal{X}} k(x, x') f(x') dx$$

En concordancia con la perspectiva que se desea abarcar, se denotará como kernel a toda función $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ en el espacio de características¹. Dentro de tal clase de funciones, son de importancia a aquellas capaces de “generalizar” el concepto de *producto interno*, el **Teorema de Mercer** permite identificar tal subconjunto.

THEOREM 3.1 (Teorema de Mercer). *Sea (\mathcal{X}, μ) un espacio de medida finita y $k \in L_{\infty}(\mathcal{X}^2)$ una función real y simétrica, tal que el operador integral:*

$$(3.2) \quad \begin{aligned} T_k : L_2(\mathcal{X}) &\rightarrow L_2(\mathcal{X}) \\ (T_k f)(x) &:= \int_{\mathcal{X}} k(x, x') f(x') d\mu(x') \end{aligned}$$

es definido positivo, es decir, para toda $f \in L_2(\mathcal{X})$ se cumple:

$$(3.3) \quad \int_{\mathcal{X}^2} k(x, x') f(x) f(x') d\mu(x) d\mu(x') \geq 0$$

NicoCaro:
 Agregar apendice sobre Teoria de Medida + integrar respecto a una medida learning with Kernels

¹El codominio del kernel k no tiene por que restringirse a los reales, para ciertas aplicaciones puede ser conviene utilizar los complejos. En general las propiedades de interés se preservan en ambos cuerpos, por lo que por simplicidad se considera solo el caso real en esta monografía.

Si $\psi_j \in L_2(\mathcal{X})$ son las funciones propias ortonormales de T_f asociadas a los valores propios $\lambda_j > 0$, ordenados de manera decreciente. Entonces,

- (1) $(\lambda_j)_j \in l_1$
- (2) $k(x, x') = \sum_{j=1}^{N_{\mathcal{H}}} \lambda_j \psi_j(x) \psi_j(x')$ se cumple para casi todos los elementos $x, x' \in \mathcal{X}$. Además $N_{\mathcal{H}} \in \mathbb{N}$ o $N_{\mathcal{H}} = \infty$, en este último caso, la serie respectiva converge absoluta y uniformemente para casi todos los elementos $x, x' \in \mathcal{X}$.

NicoCaro:
Esquema de la
demostración ?

De la segunda implicación del teorema anterior, se puede deducir que $k(x, x')$ corresponde a un producto interno en $l_2^{N_{\mathcal{H}}}$ definido a través de $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$ donde:

$$\begin{aligned} \Phi : \mathcal{X} &\rightarrow l_2^{N_{\mathcal{H}}} \\ x &\mapsto (\sqrt{\lambda_j} \psi_j(x))_{j=1, \dots, N_{\mathcal{H}}} \end{aligned}$$

Para casi todo $x \in \mathcal{X}$. En este caso, Φ se interpreta como una aplicación al espacio de características. Vale destacar, que tal espacio posee producto interno, más aún, el hecho de que exista convergencia uniforme en la serie, significa que para cualquier precisión $\varepsilon > 0$, debe existir un $n \in \mathbb{N}$ tal que k puede ser aproximado como un producto interno en \mathbb{R}^n . Lo anterior se observa al notar que para casi todo $x, x' \in \mathcal{X}$, se tiene $|k(x, x') - \langle \Phi^n(x), \Phi^n(x') \rangle| < \varepsilon$, donde $\Phi^n(x) : x \mapsto (\sqrt{\lambda_1} \psi_1(x), \dots, \sqrt{\lambda_n} \psi_n(x))$. En tal contexto, se puede interpretar al espacio de características como un espacio finito dimensional dentro de cierta precisión ε . Tal observación se concreta en el siguiente teorema:

THEOREM 3.2 (Aplicación kernel de Mercer). *Si k es un kernel que cumple las condiciones del teorema (3.1), se puede construir una aplicación Φ a un espacio donde k se comporta como un producto interno:*

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle$$

para casi todo $x, x' \in \mathcal{X}$. Más aún, para todo $\varepsilon > 0$, existe una aplicación Φ_n a un espacio n -dimensional con producto interno tal que

$$|k(x, x') - \langle \Phi_n(x), \Phi_n(x') \rangle| < \varepsilon$$

para casi todo $x, x' \in \mathcal{X}$.

La construcción que ofrece el **teorema de Mercer** permite bajo ciertas condiciones, obtener una representación del espacio inicial \mathcal{X} en un espacio con producto interno de alta dimensionalidad (dentro de cierta precisión ε). Esta construcción se obtiene a través de una descomposición espectral del operador T_k asociado al kernel k .

Al observar la representación del kernel en función de la descomposición espectral de T_k y reemplazar en la definición de positividad de este último, se obtiene:

$$(3.4) \quad \int_{\mathcal{X}^2} k(x, x') f(x) f(x') d\mu(x) d\mu(x') = \sum_j^{N_{\mathcal{H}}} \lambda_j \left[\int_{\mathcal{X}} \psi_j(x) f(x) d\mu(x) \right]^2 \geq 0$$

$\forall f \in L_2(\mathcal{X})$

Debido a las propiedades geometricas y resultados de aproximación que ofrecen los espacios con producto interno,

A Reproducing Kernel Hilbert Space (RKHS) is first of all a Hilbert space, that is, the most natural extension of the mathematical model for the actual space where

everyday life takes place (the Euclidean space \mathbb{R}^3) . When studying elements of some abstract set S it is convenient to consider them as elements of some other set S' on which is already defined a structure relevant to the problem to be treated . It can be for instance an order structure, a vector structure, a metric structure or a mixing of algebraic and topological structures. For this we need an "imbedding theorem" or a "representation theorem" . Through this kind of theorem the study of elements of S is transferred to their "representers" in S' and can be carried out using the structure on S' . For their richness and simplicity Hilbert spaces are introduced as often as possible when a vector structure and an inner product can be exploited. They provide powerful mathematical tools and geometric concepts on which our intuition can rest. The phrase "RKHS method" is generic to name a method based on the embedding of the abstract set S into some RKHS S'

CHAPTER 4

Procesos Gaussianos

El sobreajuste es al modelo
Lo que los cocodrilos a los ángeles.

Introducción

En este capítulo, se estudian los *procesos gaussianos* como herramienta de modelación no paramétrica en el contexto de aprendizaje de máquinas supervisado. En las siguientes secciones, se busca por tanto, definir los conceptos básicos referentes a dicha metodología para posteriormente estudiar sus propiedades y variantes.

4.1. Procesos Gaussianos

NicoCaro:
Poner algo profundo por el estilo. (?)


CHAPTER 5

Redes Neuronales

Qué es el *deep learning*:

Un comerciante en datos y códigos?

Un sacerdote que no cree en nada?



NicoCaro:
Poner algo profundo por el estilo. (?)

Bibliography

- [A] T. Aoki, *Calcul exponentiel des opérateurs microdifférentiels d'ordre infini*. I, Ann. Inst. Fourier (Grenoble) **33** (1983), 227–250.
- [B] R. Brown, *On a conjecture of Dirichlet*, Amer. Math. Soc., Providence, RI, 1993.
- [D] R. A. DeVore, *Approximation of functions*, Proc. Sympos. Appl. Math., vol. 36, Amer. Math. Soc., Providence, RI, 1986, pp. 34–56.

Index

- Absorbing barrier, 4
- Adjoint partial differential operator, 20
- A -harmonic function, 16, 182
- A^* -harmonic function, 182
- Boundary condition, 20, 22
 - Dirichlet, 15
 - Neumann, 16
- Boundary value problem
 - the first, 16
 - the second, 16
 - the third, 16
- Bounded set, 19
- Diffusion
 - coefficient, 1
 - equation, 3, 23
- Dirichlet
 - boundary condition, 15
 - boundary value problem, 16
- Elliptic
 - boundary value problem, 14, 158
 - partial differential equation, 14
 - partial differential operator, 19
- Fick's law, 1
- Flux, 1
- Formally adjoint partial differential operator, 20
- Fundamental solution
 - conceptual explanation, 12
 - general definition, 23
 - temporally homogeneous case, 64, 112
- Genuine solution, 196
- Green function, 156
- Green's formula, 21
- Harnack theorems
 - first theorem, 185
 - inequality, 186
 - lemma, 186
 - second theorem, 187
 - third theorem, 187
- Helmholtz decomposition, 214
- Hilbert-Schmidt expansion theorem, 120
- Initial-boundary value problem, 22
- Initial condition, 22
- Invariant measure (for the fundamental solution), 167
- Maximum principle
 - for A -harmonic functions, 183
 - for parabolic differential equations, 65
 - strong, 83
- Neumann
 - boundary condition, 16
 - boundary value problem, 16
 - function, 179
- One-parameter semigroup, 113
- Parabolic initial-boundary value problem, 22
- Partial differential equation
 - of elliptic type, 14
 - of parabolic type, 22
- Positive definite kernel, 121
- Reflecting barrier, 4
- Regular (set), 19
- Removable isolated singularity, 191
- Robin problem, 16
- Semigroup property (of fundamental solution), 64, 113
- Separation of variables, 131
- Solenoidal (vector field), 209
- Strong maximum principle, 83
- Symmetry (of fundamental solution), 64, 112
- Temporally homogeneous, 111
- Vector field with potential, 209
- Weak solution
 - of elliptic equations, 195
 - of parabolic equation, 196

associated with a boundary condition,
204