Monografía sobre Aprendizaje Automático

Nicolás Caro

Departamento de Ingeniería Matemática, Universidad de Chile, Santiago.

ncaro@dim.uchile.cl

August 27, 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
3.3. Espacios de Hilbert con kernel reproductor - RKHS 13

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 LaTeX 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

CHAPTER 1

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 capitulo se basa en la búsqueda de una representación compacta, para distribuciones de probabilidad conjunta de la forma $p(\boldsymbol{x}|\boldsymbol{\theta})$. 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 $p(\boldsymbol{x}) = p(x_1, x_2, \dots x_v)$ se puede representar de la forma:

$$p(\boldsymbol{x}) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \dots p(x_v \mid x_1, x_2, \dots, x_{v-1}) \tag{1.1}$$

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 $x_{t+1} \perp x_1, \dots, x_{t-1} \mid x_t$. 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:

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

NicoCaro: propiedades básicas del calculo 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.

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

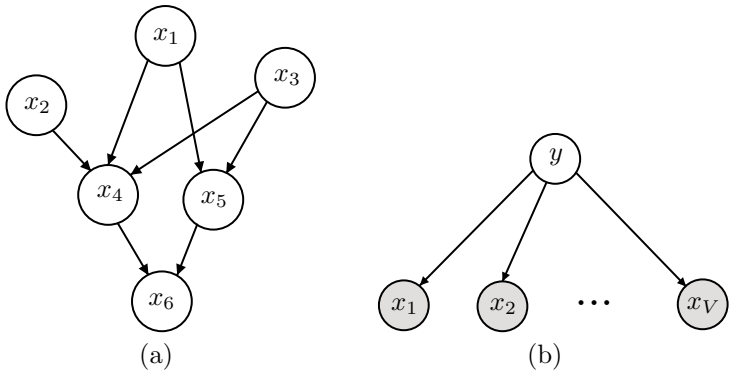

FIGURE 1. (a) Ejemplo de modelo gráfico dirigido. (b) Relaciones de dependencia condicional en el clasificador naive Bayes como un modelo gráfico dirigido, las variables aleatorias observadas se denotan por nodos grises.

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 ingles). 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:

NicoCaro: lema de ordenamiento topológico para DAG's en apéndice.

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

Es decir, un nodo x_s es independiente de aquellos predecesores, menores en orden topológico, a sus padres $\boldsymbol{x}_{pred(s)\setminus pa(s)}$, dados sus nodos padres $\boldsymbol{x}_{pa(s)}$. 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:

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

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

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

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 $s \to t$ 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.

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 $\boldsymbol{x} = (x_1, \ldots, x_V)$ en C clases. Es posible modelar las variables de decisión x_t como condicionalmente independientes dada la categoría de clasificación:

$$x_i \perp x_j \ | y = c, \ i \neq j \tag{1.6}$$

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

$$p(\boldsymbol{x} \ | y = c) = \prod_{t=1}^{V} p(x_t \ | y = c) \tag{1.7}$$

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 $\boldsymbol{w}$ y los datos observados $\boldsymbol{y} = (y_1, \ldots, y_N)^T$. Adicionalmente, se parametriza el ruido del modelo a través de σ_ε^2 y la varianza de la distribución a priori [2] de $\boldsymbol{w}$ por σ_w^2. Finalmente, los datos de entrada se denotan por $\boldsymbol{x} = (x_1, \ldots, x_N)^T$.

NicoCaro: añadir intro significativa

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

$$p(\boldsymbol{y}, \boldsymbol{w}) = p(\boldsymbol{w}) \prod_{i=1}^{N} p(y_i|\boldsymbol{w}) \tag{1.8}$$

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:

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

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

[2] Considerándose una distribución a priori, gaussiana y esférica sobre $\boldsymbol{w}$.

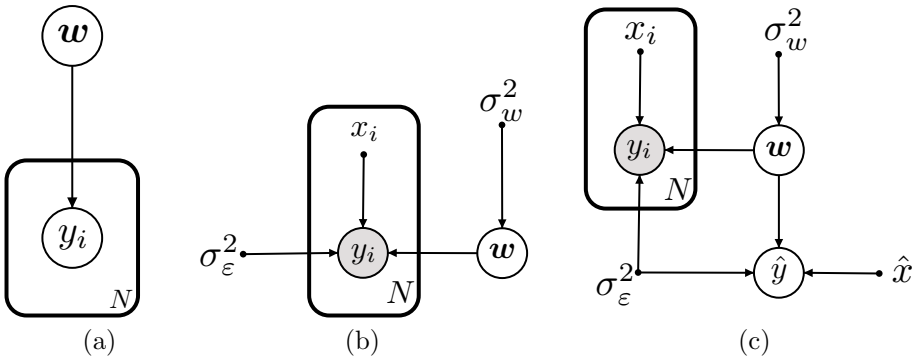

FIGURE 2. Modelo grafico dirigido para regresión polinomial usando notación de placas (o *plates*). En (a) se muestra el grafo correspondiente a (1.8). En (b) se añaden los parámetros deterministas y las variables aleatorias observadas. En (c) se añaden datos de entrada y predicciones.

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:

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

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

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

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:

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

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

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

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):

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

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

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

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

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

Es posible entonces, encontrar las componentes de $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}] = (\mathbb{E}[x_1], \ldots, \mathbb{E}[x_V])^T$ 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 $\boldsymbol{\Sigma}_{st}$ de la matriz de covarianza, observando:

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

De donde al igual que en (1.16), se calculan los elementos de $\boldsymbol{\Sigma}$ 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:

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

Donde $\boldsymbol{W}_{ts}$ 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

NicoCaro: terminar intro

En un grafo G, se escribe $x_i \perp_G x_j | x_k$ 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.

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 $p(\boldsymbol{x})$, donde $\boldsymbol{x} = (x_1, ..., x_V)^T$, posee al menos un diagrama de independencia. En efecto, si se considera un grafo G con nodos $\mathcal{V} = \{x_1, ..., x_V\}$ 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 esta *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, $s \to e \to t$, donde $e \in E$.
(2) P contiene una estructura $s \leftarrow e \to t$, donde $e \in E$.
(3) P contiene una estructura $s \to e \leftarrow t$, donde $e \notin E$ o e **no** es descendiente de algún elemento de E.

Se dice que un conjunto de nodos A esta 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 esta 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 $x \to y \to z$ 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:

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

- Sea la estructura $x \leftarrow y \to z$, 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 $x \to y \leftarrow z$, 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 si, 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:

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

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.

NicoCaro: revisar demostración Koller, Friedman 2009

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:

$$mb(t) := ch(t) \cup pa(t) \cup copa(t) \tag{1.20}$$

Donde $ch(t)$ son los nodos hijos de x_t, de manera análoga $pa(t)$ son nodos padres y $copa(t)$ sus copadres [4]. 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 unicamente en los nodos padres), sin embargo, al definir $\boldsymbol{x}_{-t}$ como el conjunto de nodos distintos a x_t, es posible observar que la probabilidad conjunta adquiere la forma $p(\boldsymbol{x}) = p(x_t, \boldsymbol{x}_{-t})$. De donde, al marginalizar sobre el nodo x_t, se obtiene que $p(\boldsymbol{x}_{-t})$ 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 $p(x_t|\boldsymbol{x}_{-t}) = p(\boldsymbol{x})/p(\boldsymbol{x}_{-t})$ solo se podrán encontrar probabilidades condicionales donde x_t sea el argumento, lo que expresa con $p(x_t|\boldsymbol{x}_{pa(t)})$, o donde sea variable condicionante, es decir, sea padre o copadre de algún otro nodo. Se deduce:

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

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

[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(a) 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.

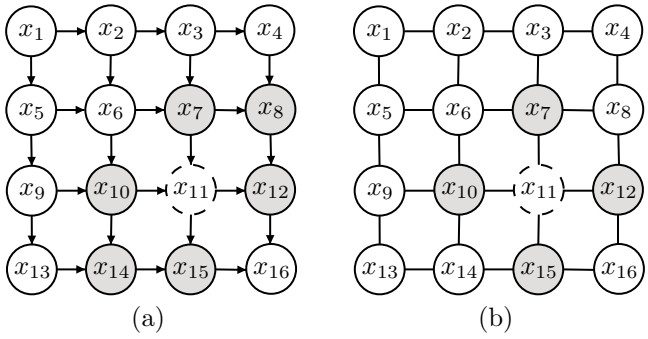

FIGURE 3. Modelo gráfico dirigido (izquierda) contrapuesto a uno no dirigido (derecha). En (a) el markov blanket de x_{11} añade a x_8 y x_{14} a sus vecinos inmediatos, mientras excluye a x_6 y x_16 en aplicaciones basadas en imágenes, este comportamiento puede no ser el deseado. Por otra parte en (b) 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.

NicoCaro:
Poner algo profundo por el estilo(?)

FT:
estoy de acuerdo

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 $\boldsymbol{x}^s \sim p(\boldsymbol{x}|\mathcal{D})$ 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

CHAPTER 3

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:

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

En concordancia con la perspectiva que se desea abarcar, se denotará como kernel a toda función $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ en el espacio de características [1]. Dentro de tal clase de funciones, son de importancia a aquellas capaces de "generalizar" el concepto de *porducto interno*, el **Teorema de Mercer** permite identificar tal subconjunto.

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

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

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

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

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

[1] El 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 $x, x' \in \mathcal{X}$.*

La segunda implicación del teorema anterior, permite construir una aplicación $\Phi : \mathcal{X} \to l_2^{N_\mathcal{H}}$ tal que $x \mapsto (\sqrt{\lambda_j}\psi_j(x))_{j=1,\dots,N_\mathcal{H}}$, es decir, a cada elemento de $\mathcal{X}$ se le asocia una transformación por medio de las funciones propias asociadas a $k(\cdot,\cdot)$ al espacio $l_2^{N_\mathcal{H}}$. Este último espacio esta provisto de producto interno, por lo que es posible expresar $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{l_2^{N_\mathcal{H}}}$ para casi todo $x \in \mathcal{X}$. Esto permite interpretar a Φ como una aplicación a un espacio de *características*, más aún, en este espacio $k(\cdot,\cdot)$ actúa como un producto interno, esto 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 convergencia uniforme en el teorema anterior implica 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. Esta característica es una parte esencial en el aprendizaje con kernels y dará lugar a la técnica conocida como *truco del kernel*.

La condición de simetría para $k(\cdot,\cdot)$, necesaria en el teorema (3.1), permite dotar de esta propiedad al producto interno definido en (3.2). Por otra parte, es necesario caracterizar la positividad de $k(\cdot,\cdot)$ de manera tal que se pueda contextualizar dentro del aprendizaje de máquinas, para ello se utiliza la noción de matriz de **Gram**.

semestre/Seminario AGW - GP/Monografia-ML-master/toLatex/3.png

Definición 3.1.2 Dado un conjunto de puntos $X_n = \{x_i \mid i = 1, ..., n\}$ y un kernel k, se define la matriz de Gram (*Gram matrix* en inglés) a la matriz K tal que $K_{i,j} = k(x_i, x_j)$.

Esta noción, permite definir la positividad de un kernel $k(\cdot,\cdot)$ en función de las propiedades que sus matrices de Gram poseen, en este sentido, el concepto de **matriz semidefinida** juega un rol fundamental.

semestre/Seminario AGW - GP/Monografia-ML-master/toLatex/1.png

Definición 3.1.1 Una matriz real K de $n \times n$ se dice que es semidefinida positiva (denotado SDP) si para todo vector $\mathbf{v} \in \mathbb{R}^n$ se cumple que $Q(\mathbf{v}) = \mathbf{v}^\top K \mathbf{v} \geq 0$. Si además se cumple que $Q(\mathbf{v}) = 0 \Leftrightarrow \mathbf{v} = 0$, entonces se dice que la matriz K es definida positiva (denotado DP).

Dentro de las propiedades de este tipo de matrices se encuentra la siguiente proposición:

semestre/Seminario AGW - GP/Monografia-ML-master/toLatex/2.png

Proposición 3.1.1 Una matriz simétrica es semidefinida positiva si y solo si todos sus valores propios son no negativos.

Finalmente, se podrá caracterizar la positividad de un kernel $k(\cdot, \cdot)$ mediante la siguiente proposición:

semestre/Seminario AGW - GP/Monografia-ML-master/toLatex/4.png

Proposición 3.1.2 Un kernel k es semidefinido positivo si y sólo si toda matriz de Gram K generada con k es semidefinida positiva. En ese caso, se dice que k es una función de covarianza.

En particular, esto significa que cualquier función simétrica y semidefinida positiva, acepta una descomposición por medio del teorema de Mercer y por tanto induce una transformación Φ a un espacio de características (posiblemente de dimensión infinita, $N_{\mathcal{H}} = \infty$) donde actúa como un producto interno. Si por otra parte, se posee una transformación $\Phi : \mathcal{X} \to \mathcal{H}$, donde $\mathcal{H}$ es un espacio con producto interno, es posible construir un kernel por medio de $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$. Este es por definición, simétrico (lo hereda del producto interno) y además cumple:

semestre/Seminario AGW - GP/Monografia-ML-master/toLatex/5.png

This can be seen by noting that for all $c_i \in \mathbb{R}, x_i \in \mathcal{X}, i = 1, \ldots, m$, we have

$$\sum_{i,j} c_i c_j k(x_i, x_j) = \left\langle \sum_i c_i \Phi(x_i), \sum_j c_j \Phi(x_j) \right\rangle = \left\| \sum_i c_i \Phi(x_i) \right\|^2 \geq 0,$$

Por tanto será semidefindo positivo y finalmente un kernel que cumple las condiciones del teorema (3.1). Esto permite modelar kernels por medio de transformaciones conocidas, a la vez que permite asumir la existencia de una transformación dado un kernel.

3.3. Espacios de Hilbert con kernel reproductor - RKHS

Hasta este punto, se puede entender un kernel como una función semidefinida positiva y simétrica compatible con una representación, ya sea implícita o explicita, de un espacio inicial $\mathcal{X}$ en un espacio con producto interno (o *pre-Hilbertiano*). El propósito de esta sección corresponde a estudiar las características de este espacio

con el fin de obtener herramientas para el uso de kernels en tareas de aproximación de funciones.

semestre/Seminario AGW - GP/Monografia-ML-master/toLatex/6.png

Definition 2.9 (Reproducing Kernel Hilbert Space) *Let $\mathcal{X}$ be a nonempty set (often called the index set) and by $\mathcal{H}$ a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. Then $\mathcal{H}$ is called a reproducing kernel Hilbert space endowed with the dot product $\langle \cdot, \cdot \rangle$ (and the norm $\|f\| := \sqrt{\langle f, f \rangle}$) if there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the following properties.*

1. *k has the reproducing property*[3]

$$\langle f, k(x, \cdot) \rangle = f(x) \text{ for all } f \in \mathcal{H}; \tag{2.34}$$

in particular,

$$\langle k(x, \cdot), k(x', \cdot) \rangle = k(x, x'). \tag{2.35}$$

2. *k spans $\mathcal{H}$, i.e. $\mathcal{H} = \overline{\text{span}\{k(x, \cdot) | x \in \mathcal{X}\}}$ where $\overline{X}$ denotes the completion of the set X (cf. Appendix B).*

Por definición (punto 2), un RKHS es un espacio formado por combinaciones lineales de funciones de la forma $k_x(\cdot)$ para todo $x \in \mathcal{X}$ y sus funciones limite (clausura). Esto último quiere decir, que toda función $f \in H$ puede ser aproximada a través de sucesiones de combinaciones lineales de elementos de la forma $k_x(\cdot)$, que claramente están unicamente determinados por el kernel $k(\cdot, \cdot)$.

Por su parte, la propiedad de reproducción (punto 1), se puede comprender al observar que para todo $f \in \mathcal{H}$, espacio RKHS correspondiente al kernel $k(\cdot, \cdot)$, es posible escribir f como:

$$f(\cdot) = \sum_{i=1}^{\infty} \alpha_i k(x_i, \cdot) \tag{3.4}$$

Que por el teorema de Mercer verifica:

$$f(x) = \sum_{i=1}^{\infty} \alpha_i \sum_{j=1}^{N_{\mathcal{H}}} \lambda_j \psi_j(x_i) \psi_j(x) \tag{3.5}$$

CHAPTER 4

Procesos Gaussianos

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

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

Introducción

En este capitulo, se estudian los *procesos gaussianos* como herramienta de modelación no paramétrica en el contexto de aprendizaje de maquinas 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

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 microdifferentiels 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
  conceptional 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

Helmhotz 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