

Monografía sobre Aprendizaje Automático

Nicolás Caro

DEPARTAMENTO DE INGENIERIA MATEMATICA, UNIVERSIDAD DE CHILE, SANTIAGO.

Email address: `ncaro@dim.uchile.cl`

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
1.3. Modelos gráficos no dirigidos	7
1.4. Inferencia exacta en modelos gráficos	7
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
Chapter 4. Procesos Gaussianos	13
Chapter 5. Redes Neuronales	15
Bibliography	17
Index	19

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

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.

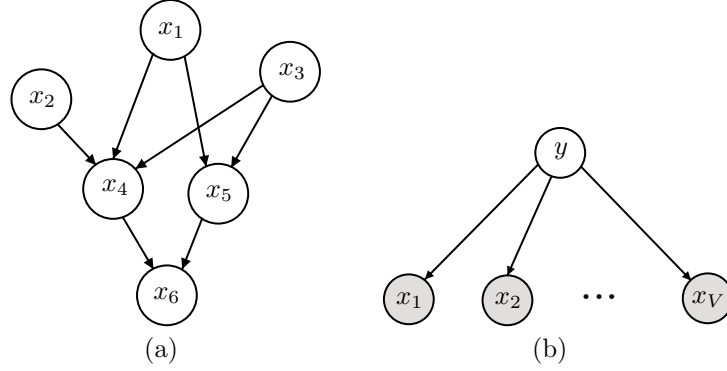


FIGURE 1. (a) Ejemplo de modelo gráfico dirigido. (b) Relaciones de dependencia condicional en el clasificador naive Bayes como un modelo gráfico dirigido, las variables aleatorias observadas se denotan por nodos grises.

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:

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

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 $\mathbf{x} = (x_1, \dots, x_N)^T$.

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

NicoCaro:
añadir intro significativa

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

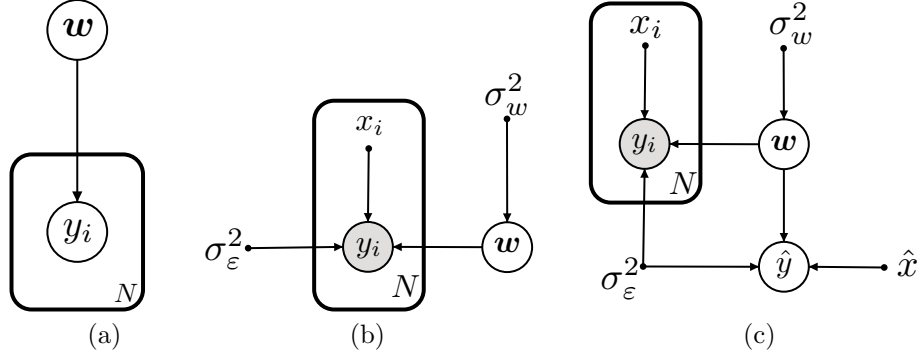


FIGURE 2. Modelo grafico dirigido para regresión polinomial usando notación de placas (o *plates*). En (a) se muestra el grafo correspondiente a (1.8). En (b) se añaden los parámetros deterministas y las variables aleatorias observadas. En (c) se añaden datos de entrada y predicciones.

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 $\boldsymbol{\Sigma}_{st}$ 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 $\boldsymbol{\Sigma}$ 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:

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

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 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.

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 $G' = G$.

A continuación se estudian las características de los modelos gráficos dirigidos que permiten determinar cuando existe independencia condicional en sus nodos.

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:

$$(1.19) \quad \mathbf{x}_A \perp_G \mathbf{x}_B | \mathbf{x}_E \iff A \text{ está d-separado de } B \text{ dado } E$$

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:

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

- 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 requiera en este tipo de estructuras, que no existan nodos, ni descendientes de nodos de la familia condicionante E .

Si bien en el último caso, no se cumplía $x \perp z|y$, si se tiene $p(x, z) = p(x)p(z)$, es decir, x y z son marginalmente independientes, sin embargo, al condicionar

ambos nodos por un nodo hijo y , se vuelven dependientes, este efecto se denomina **paradoja de Berkson**.

1.3. Modelos gráficos no dirigidos

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

Inferencia Markov chain Monte Carlo

Inferencia variacional

NicoCaro:
Poner algo profundo por el estilo(?)

FT:
estoy de acuerdo

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 pro-
fundo por el es-
tilo. (?)

CHAPTER 4

Procesos Gaussianos

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

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