EST-25134: Aprendizaje Estadístico

Profesor: Alfredo Garbuno Iñigo — Primavera, 2022 — Máquinas de soporte vectorial.
Objetivo: Estudiaremos una de las técnicas clásicas de modelado predictivo por medio de máquinas de soporte vectorial (`SVM`). Estos modelos encuentran relaciones no lineales y, además, hacen enfásis en el diseño de características (atributos) para tareas de modelado. En particular, son una alternativa viable para datos pequeños donde se tiene mucho conocimiento de dominio para el diseño de mecanismos de ajuste. *Disclaimer*: Todas las figuras en esta sección fueron tomadas de [3].
Lectura recomendada: Capítulo 9 de [3], Capítulo 4 y 12 de [2].

1. INTRODUCCIÓN

Las máquinas de soporte vectorial (`SVM`) son modelos no lineales que buscan resolver problemas predictivos. Por simplicidad, estudiaremos `SVM` en el contexto de clasificación.

La forma en que operan las `SVM` es a través de buscar clasificar por un `separación con un hiperplano` en el espacio de atributos.

Si los datos no nos permiten realizar dicha separación entonces nos volcamos a relajar los supuestos del modelo:

1. El concepto de `separación`;
2. El espacio de atributos.

2. SEPARACIÓN CON UN HIPERPLANO

- Un hiperplano de dimensión p es un subespacio de dimensiones $p-1$.
- La ecuación que define a un hiperplano tiene la forma

$$0 = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{1}$$

- En dos dimensiones es una línea.
- Si $\beta_0 = 0$ entonces el hiperplano atraviesa el origen.
- El vector de coeficientes $(\beta_1, \ldots, \beta_p)$ se llama el vector `normal` del hiperplano.

2.1. Para problemas de clasificación

- Nota que si definimos $f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, entonces $f(X) > 0$ para X de un lado del hiperplano. Si $f(X) < 0$ entonces X se encuentra del otro lado.
- Para los datos en Fig. 1, si los puntos azules son aquellos donde $Y_i = 1$ y los morados son aquellos que tienen $Y_i = -1$, entonces

$$Y_i \cdot f(X_i) > 0\,, \qquad \forall i\,. \tag{2}$$

- El caso $f(X) = 0$ define al `hiperplano separador`.

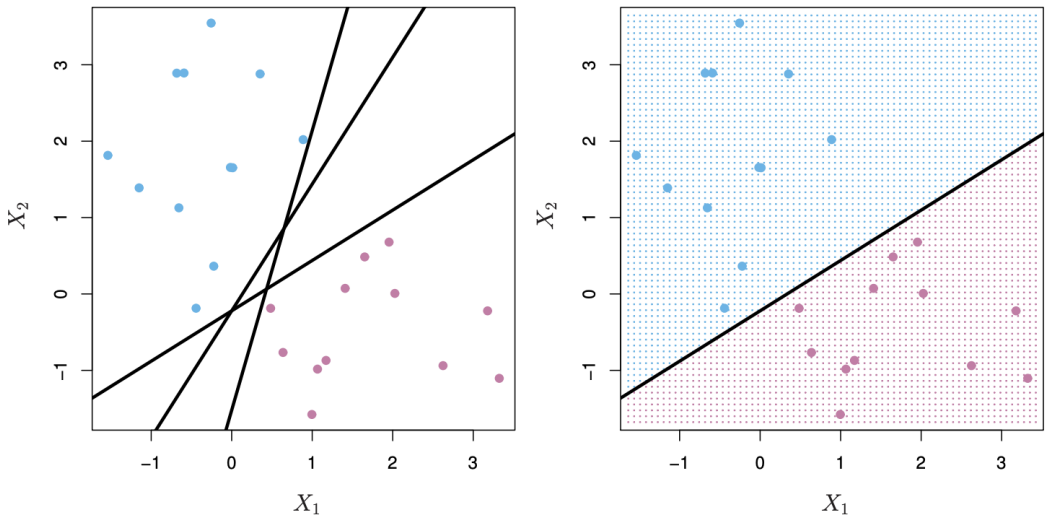
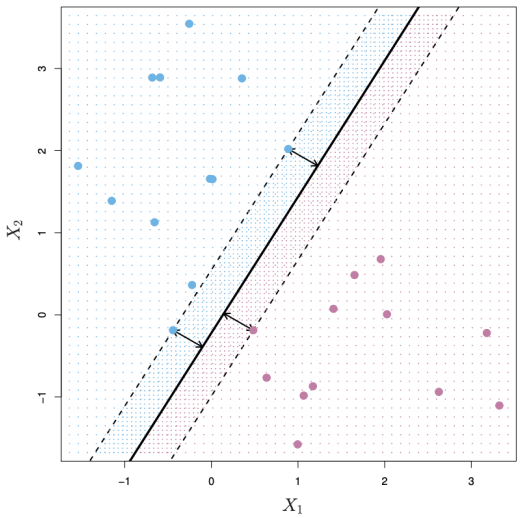


FIGURA 1. *Imagen tomada de [3].*

2.2. Máxima separación

En Fig. 1 vemos que existe una infinidad de planos que pueden separar nuestro conjunto de datos. Nuestra noción del `mejor` separador es aquel que `maximice el margen` de separación.



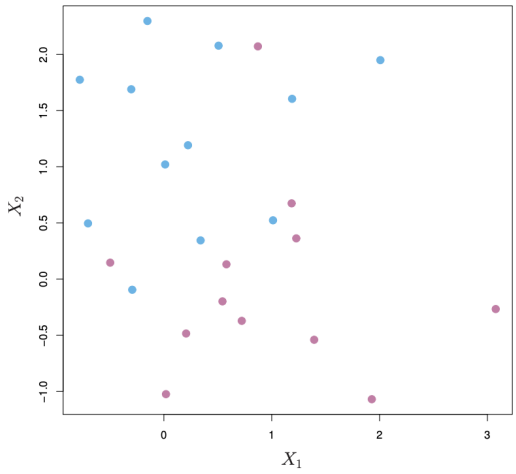
Esto lo escribimos como la solución al problema de optimización con restricciones

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} M$$
$$\text{sujeto a } \sum_{j=1}^{p} \beta_j^2 = 1\,,$$
$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M\,, \quad \forall i\,.$$

El problema de optimización descrito se denomina `formulación primal` y se puede rescribir en términos de un problema de optimización cuadrático por medio de la `formulación dual`. Esto permite la resolución del problema de optimización por medio de herramientas de optimización convexa para problemas cuadráticos.

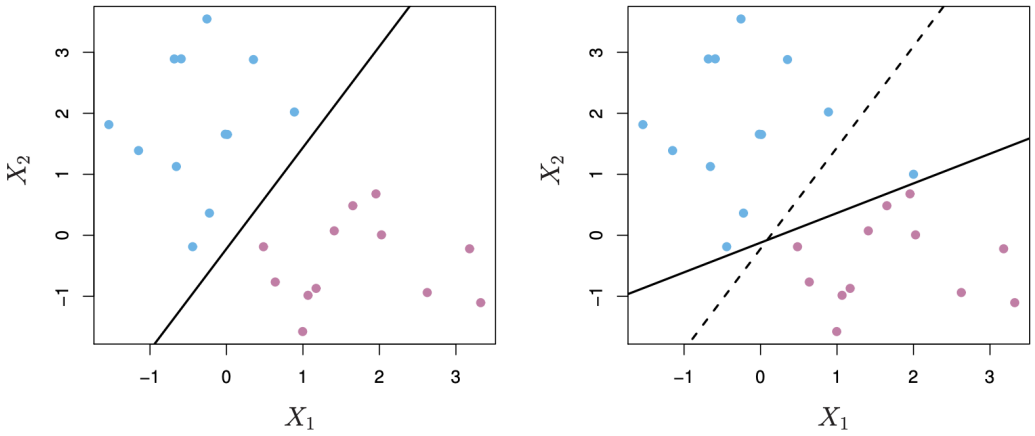
2.3. Datos no separables

- Usualmente sucede esto en la práctica (a menos que $n < p$).



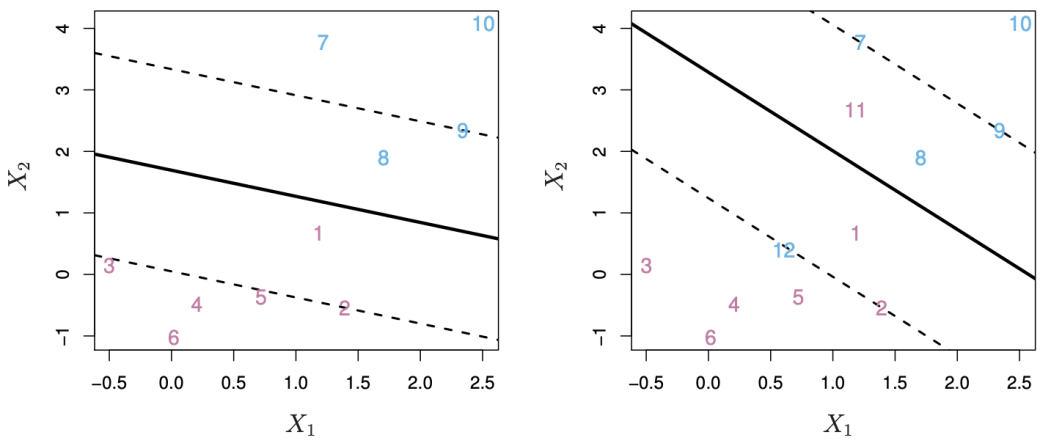
2.4. Ruido en las mediciones

Los datos a veces son separables, pero el ruido en las observaciones puede hacer que un clasificador por margen máximo tenga generalización deficiente.



3. CLASIFICADOR BASADO EN VECTORES DE SOPORTE

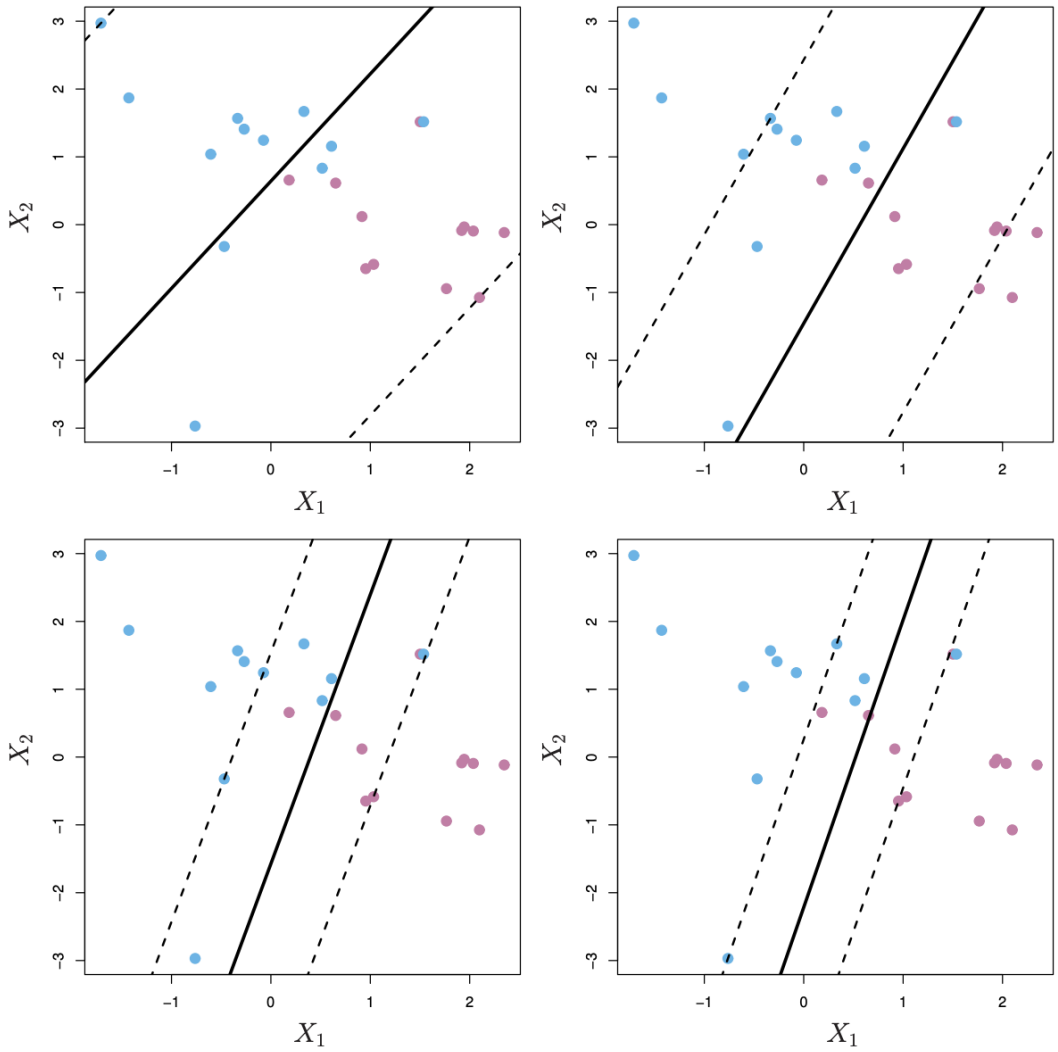
Se pueden relajar las restricciones del clasificador para incorporar un `margen suave`.



3.1. Formulación

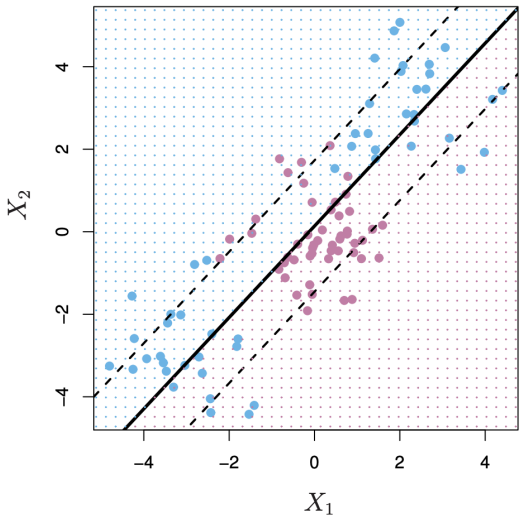
$$\max_{\beta_0,\beta_1,\ldots,\beta_p} M$$
$$\text{sujeto a } \sum_{j=1}^{p} \beta_j^2 = 1\,,$$
$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M(1-\epsilon_i)\,, \quad \forall i\,,$$
$$\epsilon_i \geq 0\,, \qquad \sum_{i=1}^{n} \epsilon_i \leq C\,.$$

El término C dicta cuántos datos mal clasificados estamos dispuestos a cometer. Las variables ϵ_i reciben el nombre de variables de `holgura`.



4. SEPARACIÓN LINEAL

En algunas situaciones la separación lineal no será suficiente.



4.1. Ingeniería de características

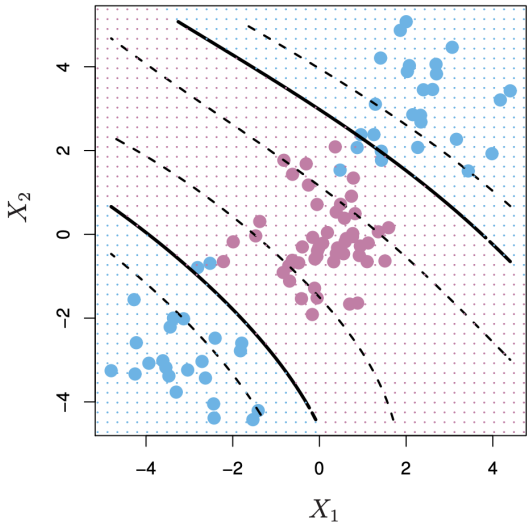
- Buscamos una expansión de los atributos por medio de transformaciones

$$X_1^2, X_2^2, X_1X_2, \dots \tag{3}$$

  lo cual permite expandir el espacio $\mathbb{R}^p \to \mathbb{R}^M$ con $M > p$.
- Ajustamos un clasificador por vectores de soporte en el espacio expandido.
- Esto nos ayuda a incorporar decisión de separación no lineales en el espacio original de atributos.

4.2. Solución con polinomios cúbicos

- Si utilizamos una expansión con polinomios cúbicos, pasamos de 2 dimensiones a 9.
- La separación lineal la logramos en este nuevo espacio.



5. SEPARACIONES NO LINEALES Y KERNELS

- Los polinomios (especialmente en varias dimensiones) pueden tener un comportamiento oscilatorio muy fuerte y sobre-ajustar sin cuidado.
- Hay un mecanismo matemáticamente mas elegante para introducir no-linealidades.
- Especialmente útil en problemas donde podamos usar `productos interiores`.

5.1. Formulación alterna

Vamos a reformular el problema de optimización (sin holguras)

$$\begin{aligned} &\max_{\beta_0,\beta_1,\dots,\beta_p} M \\ &\text{sujeto a: } \sum_{j=1}^{p} \beta_j^2 = 1\,, \\ &y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M\,, \quad \forall i\,. \end{aligned}$$

Si consideramos que podemos eliminar la restricción para $\|\beta\| = 1$ por medio de

$$y_i(\beta_0 + x_i^\top \beta) \geq M \cdot \|\beta\|\,, \tag{4}$$

y $\|\beta\| = 1/M$, entonces el problema de optimización lo podemos escribir como

$$\begin{aligned} &\min_{\beta_0,\beta_1,\ldots,\beta_p} \frac{1}{2}\|\beta\|^2 \\ &\text{sujeto a: } y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq 1\,, \quad \forall i\,. \end{aligned}$$

Este un problema de optimización cuadrático con restricciones lineales. Lo cual se puede resolver de manera eficiente.

Para el problema con variables de holgura

$$\begin{aligned} &\max_{\beta_0,\beta_1,\ldots,\beta_p} M \\ &\text{sujeto a: } \sum_{j=1}^{p} \beta_j^2 = 1\,, \\ &y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M(1-\epsilon_i)\,, \quad \forall i\,, \\ &\epsilon_i \geq 0\,, \qquad \sum_{i=1}^{n} \epsilon_i \leq C_{\mathsf{rest}}\,. \end{aligned}$$

Se puede reformular por medio de

$$\begin{aligned} &\min_{\beta_0,\beta_1,\ldots,\beta_p} \frac{1}{2}\|\beta\|^2 + C_{\mathsf{reg}} \sum_{i=1}^{n} \epsilon_i \\ &\text{sujeto a:} \\ &y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq 1 - \epsilon_i\,, \quad \forall i\,, \\ &\epsilon_i \geq 0\,. \end{aligned}$$

5.2. Productos interiores y vectores de soporte

- Recordemos que

$$\langle x_i, x_{i'} \rangle = x_i^\top x_{i'} = \sum_{j=1}^{p} x_{ij} x_{i'j}\,. \tag{5}$$

- El clasificador se puede expresar como

$$f(x) = \beta_0 + \sum_{i=1}^{n} y_i\, \alpha_i\, \langle x, x_i \rangle \tag{6}$$

- Necesitamos los $\binom{n}{2}$ productos interiores para poder estimar los parámetros.
- Pero, la mayoría de las α_i son cero,

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} y_i\, \alpha_i\, \langle x, x_i \rangle\,. \tag{7}$$

5.3. Kernels y máquinas de soporte vectorial

- Podríamos calcular productos interiores y construir un clasificador por medio de vectores de soporte.

- Hay algunos *kernels* que calculan lo que necesitamos. Por ejemplo,

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d , \tag{8}$$

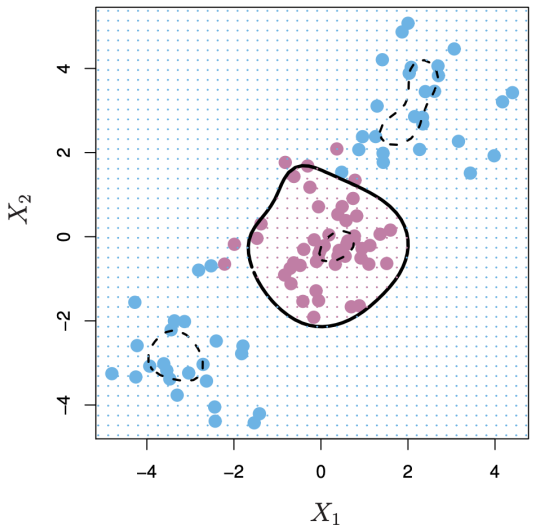
  calcula los productos internos de la expansión en polinomios.
- La solución del problema, entonces, tiene la forma

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} y_i \, \alpha_i \, K(x, x_i) \, . \tag{9}$$

5.3.1. Kernel *radial:* Consideremos el *kernel*

$$K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{p} (x_i - x_{i'})^2\right) , \tag{10}$$

lo cual nos permite ajustar superficies de decisión como la que muestra



6. CLASIFICACIÓN MULTICLASE

Una pregunta natural es cómo extender la formulación de una `SVM` para el problema de clasificación multiclase.

Tenemos dos estrategias:

1. `Una clase contra todas las anteriores`. Clasifica la clase con mayor $\hat{f}_k(x)$.
2. `Todas las posibles tareas de dos clases`. Clasifica la clase que gana la mayor de las veces.

7. COMPARACIÓN CON OTROS MODELOS

El problema de optimización se puede rescribir como

$$\min_{\beta_0, \beta_1, \ldots, \beta_p} \left\{ \sum_{i=1}^{n} \text{máx}[0, 1 - y_i f(x_i)] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} . \tag{11}$$

La función de pérdida se conoce como la función `hinge`. Es muy similar a la pérdida por entropía cruzada (regresión logística).

La formulación anterior no es muy común pero ha empezado a tener importancia pues permite formular de manera alternativa el problema de hiperplanos separadores de márgen máximo para poder resolverlo con métodos de descenso con gradiente estocástico [1].

7.1. Regresión logística ó SVM

- Cuando las clases son casi separables, preferimos `SVM`.
- Cuando no, regresión logística + Ridge es muy parecido a `SVM`.
- Si nos interesa calcular probabilidades, usamos regresión logística.
- Se pueden utilizar *kernels* con otros modelos (regresión logística o LDA) pero es mas costoso.

8. CONCLUSIONES

- Las `SVM` son modelos que pueden acomodar no linealidades para hacer modelos predictivos.
- En la práctica aún siguen siendo utilizadas pues es de las pocas alternativas para tratar relaciones no lineales entre predictores.
- Siguen siendo muy útiles en aplicaciones donde se tienen identificadas las características mas importantes para una tarea predictiva.
- Los `SVM` han permitido el desarrollo de teoría para ciertas condiciones de aprendizaje y métricas de complejidad [4].

REFERENCIAS

[1] V. Abeykoon, G. Fox, M. Kim, S. Ekanayake, S. Kamburugamuve, K. Govindarajan, P. Wickramasinghe, N. Perera, C. Widanage, A. Uyar, G. Gunduz, and S. Akkas. Stochastic gradient descent-based support vector machines training optimization on Big Data and HPC frameworks. *Concurrency and Computation: Practice and Experience*, 34(8):e6292, 2022. ISSN 1532-0634. . 9

[2] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2009. ISBN 978-0-387-84857-0 978-0-387-84858-7. . 1

[3] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. Springer US, New York, NY, 2021. 1, 2

[4] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014. ISBN 978-1-107-05713-5. 9