
quantil

Inferencia en Modelos Gráficos para detección de Tópicos (LDA)

Francisco Barreras

19 de enero de 2017

Contenido

Introducción y Contexto

Tipos de análisis en Machine Learning

- Se pueden dividir los modelos de aprendizaje de Máquinas en **problemas de clasificación** y en **predicción estructurada**.
- ¿Queremos aprender el valor de una variable real o categórica?
- ¿O aprender las probabilidades de variables aleatorias que tienen interdependencias complejas?
- A veces queremos aprender la estructura dentro de los datos.
- Podemos inferir la probabilidad de variables no observadas.

Predicción estructurada

- **Machine Translation**. No es un problema de clasificación, sino que queremos predecir cada palabra de manera que el todo tenga sentido.
- **Audio a texto** (y en general cualquier problema de canal ruidoso)
- **Diagnóstico Médico Usando Big Data** Teniendo miles de enfermedades y miles de síntomas, no es tan útil un modelo que me diga (con margen de error) si tengo una enfermedad. Quiero un modelo que me diga, dados los síntomas que tengo y que no tengo, cuál es la probabilidad de distintas enfermedades.

Motivación

Forma de Modelamiento

- Una aproximación general para que un computador **razone** sobre diversos problemas y responda preguntas.
- **Representación Declarativa:** Un modelo de cómo razonar, con algoritmos de propósito general. Separados del conocimiento y los datos en particular.
- Una principal preocupación es modelar la incertidumbre (En un sentido probabilístico).
- Casi nunca se puede descartar algo como imposible dada una observación, pero si es útil identificar **eventos improbables**.

Forma de Modelamiento

- Vamos a razonar sobre **variables aleatorias** dentro de un **sistema complejo**.
- Para esto se construye una **distribución de probabilidad conjunta**, y se busca que no sea imposible de aprender (intractable).
- Se pueden hacer muchas preguntas al modelo, por ejemplo, al observar el valor X_i como cambia la probabilidad de X_j (**probabilidad posterior**).

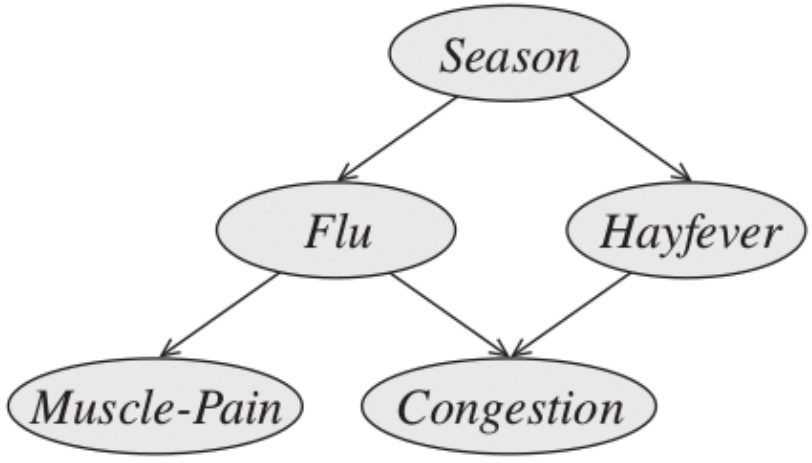
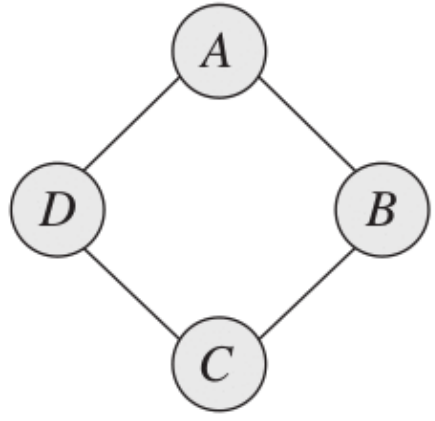
Ejemplo: Rinitis vs. Gripa

- Enfermedades no excluyentes.
- Hay una variable relevante que es la estación.
- Observamos dos síntomas: Congestión y dolor muscular.
- La probabilidad conjunta entonces debe tener $2x2x4x2x2 = 64$ valores.
- ¿Probabilidad de que un paciente tenga Gripa si es Otoño, tiene congestión pero no dolor muscular?

$$P(\text{Gripa}|E = \text{Otoño}, C = 1, DM = 0)$$

Modelos Gráficos Probabilísticos

- Especificar una distribución de 64 valores de por si puede ser retador.
- En las aplicaciones reales, puede haber cientos o miles de variables, los métodos usuales se vuelven incomputables.
- Explotando la estructura en el **sistema complejo** podemos construir una **representación compacta**.
- Esto se hace con **grafos**, donde cada **nodo** está asociado a una **variable aleatoria** y cada arista a una relación probabilística entre ellos.

	Bayesian networks	Markov networks
Graph Representation		
Independencies	$(F \perp H \mid S)$ $(C \perp S \mid F, H)$ $(M \perp H, C \mid F)$ $(M \perp C \mid F)$	$(A \perp C \mid B, D)$ $(B \perp D \mid A, C)$
Factorization	$P(S, F, H, C, M) = P(S)P(F \mid S)$ $P(H \mid S)P(C \mid F, H)P(M \mid F)$	$P(A, B, C, D) = \frac{1}{Z}\phi_1(A, B)$ $\phi_2(B, C)\phi_3(C, D)\phi_4(A, D)$

Primera perspectiva: Codificar Independencias

- El grafo informa sobre cuáles independencias condicionales existen en una distribución.
- Por ejemplo, $(X \perp Y|Z)$ lo cual se lee:

$$P(X|Y,Z) = P(X|Z)$$

- El grafo dice que un nodo es independiente de sus **no descendientes condicional en sus padres** (en BN).

Segunda Perspectiva: Factorización

- El grafo es una factorización de la distribución conjunta que tiene menor dimensionalidad.
- Por ejemplo una cadena de Markov es un modelo gráfico donde yo supongo que:

$$P(X_1, X_2, \ldots, X_t) = P(X1)P(X_2|X_1)P(X_3|X_2)\ldots P(X_t|X_{t-1})$$

- En el ejemplo de la rinitis, reduce la dimensionalidad de 63 parámetros a estimar, a sólo 17.
- Ambas aproximaciones son **equivalentes** en un sentido más profundo y específico.

Representación, Inferencia, Aprendizaje

- Una forma transparente e interpretable de **representar** Incluso si hay miles de variables (Plantillas)
- Diferentes métodos y algoritmos de cómo hacer inferencia, incluso si en algunos casos incomputables o variables faltantes. (Métodos variacionales, Gibbs Sampling, Métodos bayesianos)
- Es **aprendizaje** de máquinas. Tras especificar un modelo, se puede aprender las distribuciones e inclusive las conexiones entre variables.

Datos históricos

- Los fundamentos de la probabilidad se pueden rastrear hasta Gerolamo Cardano (S XVI) quien dio los primeros conceptos para analizar juegos de azar.
- Fermat y Pascal le dieron formalidad en el S XVII
- Bayes en el siglo XVIII mostró el llamado *Teorema de Bayes* una idea que usa la probabilidad condicional de A dado B para calcular la contrapositiva.
- La generalización de esta idea sobre relaciones complejas motiva el nombre de redes bayesianas.

Datos históricos

- Fue Gibbs (1902) el primero que representó la interacción de múltiples variables en una distribución multidimensional como un grafo. (Física Estadística).
- En genética el primero fue Sewal Wright (1921,1933) quien propuso un modelo para la herencia genética.
- En Computación, los primeros modelos eran basados en Naive Bayes (de Bombal et al. 1972; Gorry y Barnett 1968; Warner et al. 1961) para automatizar tareas complejas.
- Cayeron en desuso por la idea equivocada de que el futuro de la computación era replicar la inteligencia humana (no basada en números) y porque los supuestos de Naive Bayes eran muy restrictivos.

Definiciones y Fundamentos

Conceptos básicos

- Factor.
- Distribución de probabilidad condicional.
- Interpretación Frecuentista de Probabilidad.
- Interpretación subjetivista de Probabilidad.
- Regla de la cadena.
- Regla de Bayes.

Regla de la cadena

$$P(X_1 \ldots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \ldots P(X_n|X_{n-1}, \ldots, X_1)$$

Regla de Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Preguntas probabilísticas

- Tenemos Y la variable sobre la que se está razonando.
- Tenemos $E = e$ la instanciación de otras variables, que llamamos la evidencia.
- La pregunta más común es la probabilidad posterior:

$$P(Y|E = e)$$

- Lo cuál se puede entender como la probabilidad marginal de Y condicional a esa evidencia.

Preguntas MAP

- *Maximum A Posteriori*
- La instancia más probable de todas las variables no observadas $W = \aleph - E$
- Típicamente es diferente de los valores que maximizan cada marginal:

$$\text{MAP}(W|e) = \arg\max_{\omega} P(\omega, e)$$

- Síntomas típicos dada una enfermedad.

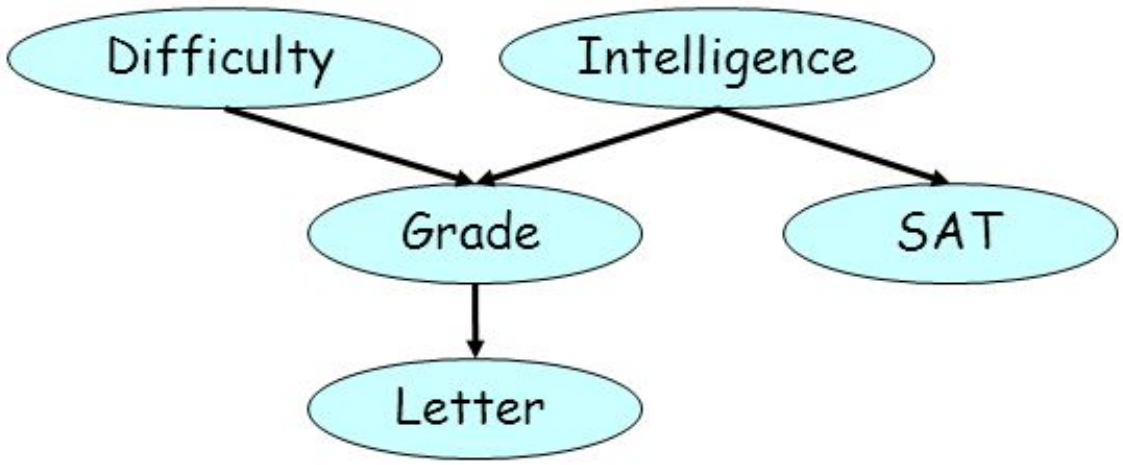
MAP marginal

- Podemos preguntar el MAP para una sola variable, marginalizando
- Si $Z = \aleph - E - Y$ entonces

$$\text{MAP}(Y|e) = \arg\max_{\omega} \sum_{Z} P(\omega, Z|e)$$

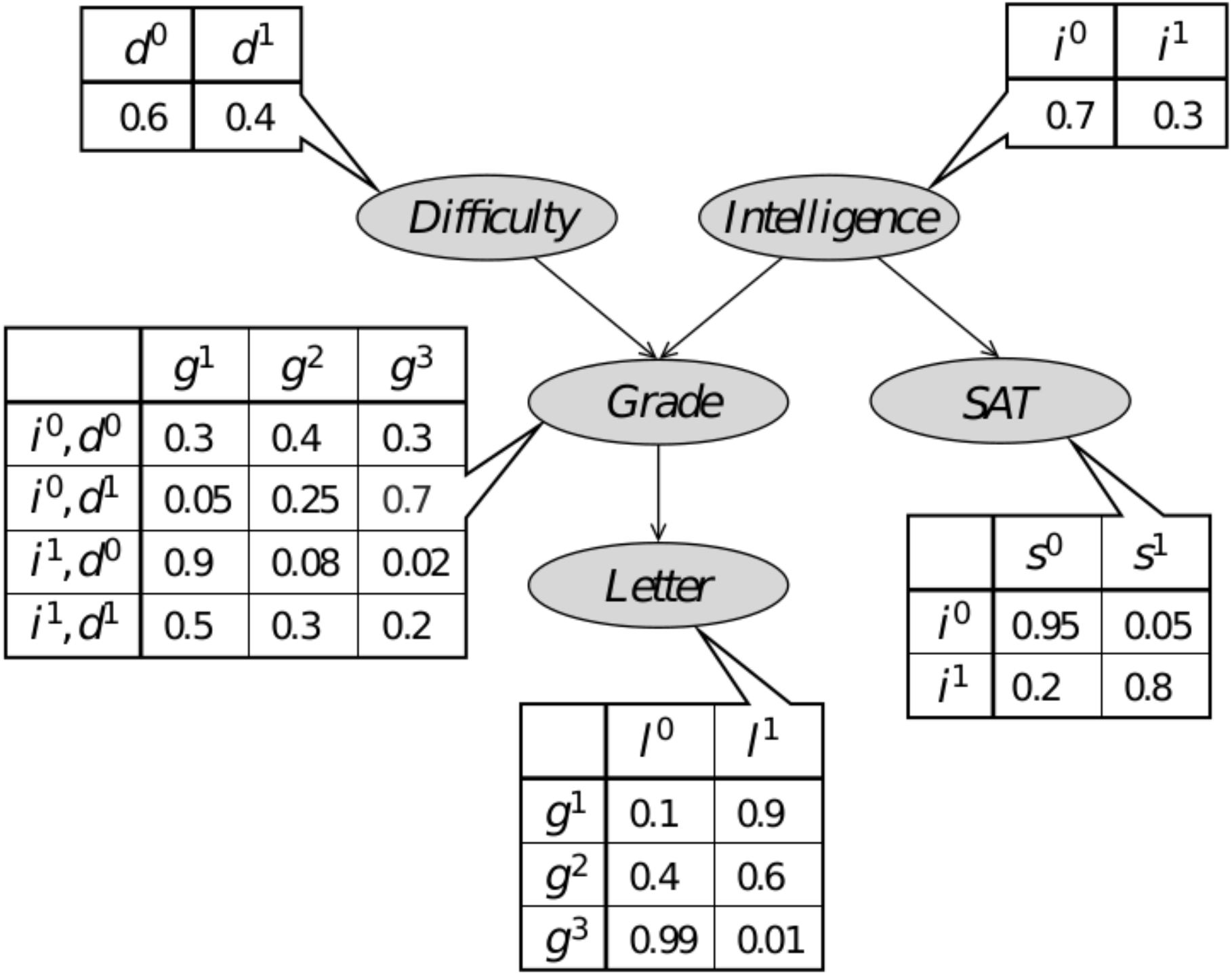
- Enfermedad más probable dados unos síntomas (clasificación?)

Razonamiento y flujos de información


Difficulty
Intelligence
Grade
SAT
Letter

Modelo del estudiante

- Vamos a tener 5 variables aleatorias:
  - I es la inteligencia del estudiante
  - D es la dificultad del curso
  - G es la nota del estudiante en el curso
  - S es el puntaje del estudiante en el Icfes
  - L es si recibe una carta de recomendación.
- Por simplicidad todas serán variables binarias.


d^0 d^1
0.6 0.4
i^0 i^1
0.7 0.3
Difficulty
Intelligence
Grade
SAT
Letter
g^1 g^2 g^3
i^0, d^0 0.3 0.4 0.3
i^0, d^1 0.05 0.25 0.7
i^1, d^0 0.9 0.08 0.02
i^1, d^1 0.5 0.3 0.2
s^0 s^1
i^0 0.95 0.05
i^1 0.2 0.8
l^0 l^1
g^1 0.1 0.9
g^2 0.4 0.6
g^3 0.99 0.01

Modelo del estudiante

- Una red bayesiana es un Grafo Dirigido Acíclico (DAG).
- La segunda cosa por definir son los modelos de **probabilidad local**.
- Con esta información podemos calcular cualquier probabilidad dentro del modelo.
- Por ejemplo:

$$\begin{aligned} P(i^1, d^0, g^2, s^1, l^0) &= P(i^1)P(d^0)P(g^2|i^1, d^0)P(s^1|i^1)P(l^0|g^2) \\ &= 0{,}3 * 0{,}6 * 0{,}08 * 0{,}8 * 0{,}4 = 0{,}004608 \end{aligned}$$

Distribución de probabilidad?

- Hasta ahora sólo hemos presentado una factorización para una supuesta distribución conjunta.
- Es necesario demostrar que en efecto es una probabilidad sobre las variables aleatorias.
- Es trivial ver que es positiva, sólo haría falta probar que:

$$\sum_{I,D,G,S,L} P(I)P(D)P(G|I,D)P(L|G)P(S|I) = 1$$

De manera más general probaremos que las factorizaciones siempre nos dan una distribución de probabilidad.

Ejemplo: Inferencia causal

- Es cuando las causas influencian nuestro posterior.
- Juan quiere saber la probabilidad de que le den una carta en el curso de modelos gráficos.
- Sin saber nada tenemos: $P(I^1) = 0{,}502$
- Si en cambio sabemos que Juan no es muy inteligente $P(I^1|i^0) = 0{,}389$.
- Pero en cambio si sabemos que el curso no era muy difícil $P(I^1|i^0, d^0) = 0{,}513$

Ejemplo: Inferencia evidencial

- Es cuando la evidencia nos cambia las creencias sobre las causas.
- Alguien en Quantil quiere saber si contrata a Juan, el cree que Juan es inteligente con probabilidad $P(i^1) = 0{,}30$
- Pero el reclutador observa que Juan sacó C en el curso, entonces ahora $P(i^1|g^C) = 0{,}079$
- Si el reclutador no tiene la nota sino una mala carta, también disminuye su probabilidad... pero menos $P(i^1|l^0) = 0{,}14$
- Y si tiene ambas sucede que $P(i^1|l^0, g^C) = 0{,}079$. esto tiene sentido?

Ejemplo: Inferencia intercausal

- La evidencia permite que una causa explique algo que le achacábamos a otra causa. (Explaining away)
- El reclutador recibe el Icfes de Juan y observa que es alto, entonces su probabilidad de ser inteligente sube dramáticamente $P(i^1|g^3, s^1) = 0{,}578$.
- Es más probable que un estudiante inteligente falle en una clase difícil a que un estudiante no inteligente saque un Icfes muy alto.
- En el fondo estamos diciendo que la probabilidad de que el curso sea difícil subió, por la observación del Icfes (**dada la mala nota**) de 0,629 a 0,761.

Ejemplo: Naive Bayes