DSC250: Advanced Data Mining

Machine Learning Basics

Zhiting Hu

Lecture 2, October 3, 2023

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Outline

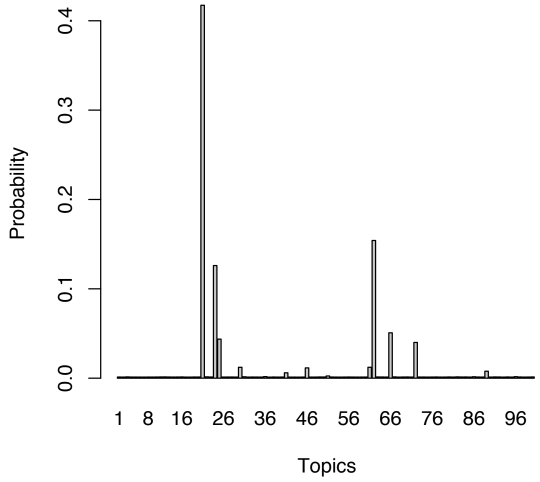
- Probability
  - Bayes' rule
  - Exponential family
  - Probabilistic graphical models
  - Entropy, KL divergence, cross entropy
- Functional derivatives
- Practice: MLE vs Maximum entropy

Probability

Why Probability?

- The world is a very uncertain place
    - “What will the weather be like today?”
    - “Will I like this movie?”
- We often can’t prove something is true, but we can still ask how likely different outcomes are or ask for the most likely explanations
- Predictions need to have associated confidence
    - Confidence -> probability
- Not all machine learning models are probabilistic
    - … but most of them have probabilistic interpretations

Example: topic modeling



"Genetics"	**"Evolution"**	**"Disease"**	**"Computers"**
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

For documents in a large collection of text, model $p(\mathrm{Word}|\mathrm{Topic})$, $p(\mathrm{Topic})$

Figure from (Blei, 2011), shows topics and top words learned automatically from reading 17,000 Science articles

[15-780, Kolter]

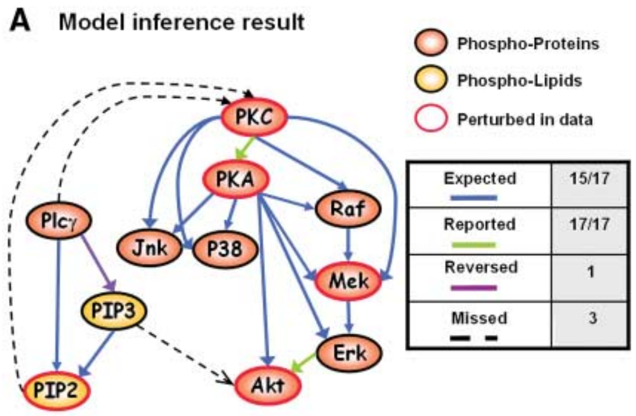
Example: image segmentation

Figure (Nowozin and Lampert, 2012) shows image segmentation problem, original image on left, where goal is to separate foreground from background

Middle figure shows a segmentation where each pixel is individually classified as belonging to foreground or background

Right figure shows a segmentation where the segmentation is inferred from a probability model over all pixels jointly (encoding probability that neighboring pixels tend to belong to the same group)

Example: modeling protein networks



In cellular modeling, can we automatically determine how the presence or absence of some proteins affects other proteins?

Figure from (Sachs et al., 2005) shows automatically inferred protein probability network, which captured most of the known interactions using data-driven methods (far less manual effort than previous methods)

Notations

- A random variable $\boldsymbol{x}$ represents outcomes or states of the world.
  - We write $p(\boldsymbol{x}_0)$ to mean Probability($\boldsymbol{x} = \boldsymbol{x}_0$)
- Sample space: the space of all possible outcomes (may be discrete, continuous, or mixed)
- $p(\boldsymbol{x})$ is the probability mass (density) function
  - Assigns a number to each point in sample space
  - Non-negative, sums (integrates) to 1
  - Intuitively: how often does $\boldsymbol{x}$ occur, how much do we believe in $\boldsymbol{x}$.

Notations

- Joint distribution $p(\boldsymbol{x}, \boldsymbol{y})$
- Conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$
  - $p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})}$
- Expectation:

$$\mathbb{E}[f(\boldsymbol{x})] = \sum_{\boldsymbol{x}} f(\boldsymbol{x})\, p(\boldsymbol{x})$$

or

$$\mathbb{E}[f(\boldsymbol{x})] = \int_{\boldsymbol{x}} f(\boldsymbol{x}) p(\boldsymbol{x}) dx$$

Rules of Probability

- Sum rule

$$p(x) = \sum_y p(x, y)$$ (Marginalize out y)

$$p(x_1) = \sum_{x_2} \sum_{x_3} \cdots \sum_{x_N} p(x_1, x_2, \ldots, x_N)$$

- Product/chain rule

$$p(x, y) = p(y \mid x) p(x)$$

$$p(x_1, \ldots, x_N) = p(x_1) p(x_2 \mid x_1) \ldots p(x_N \mid x_1, \ldots, x_{N-1})$$

Bayes' Rule

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

- This gives us a way of "reversing" conditional probabilities
- We call $p(\boldsymbol{y})$ the "prior", and $p(\boldsymbol{y}|\boldsymbol{x})$ the "posterior"
- Ex: Bayes' Rule in machine learning:
  - $\mathcal{D}$: data (evidence)
  - $\boldsymbol{\theta}$: unknown quantities, such as model parameters, predictions

Likelihood: How likely is the observed data under the particular unknown quantities $\boldsymbol{\theta}$

Posterior belief on the unknown quantities you see data $\mathcal{D}$

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

Prior belief on the unknown quantities **before** you see data $\mathcal{D}$

Independence

- Two random variables are said to be **independent** iff their joint distribution factors

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y})$$

- Two random variables are **conditionally independent** given a third if they are independent after conditioning on the third

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

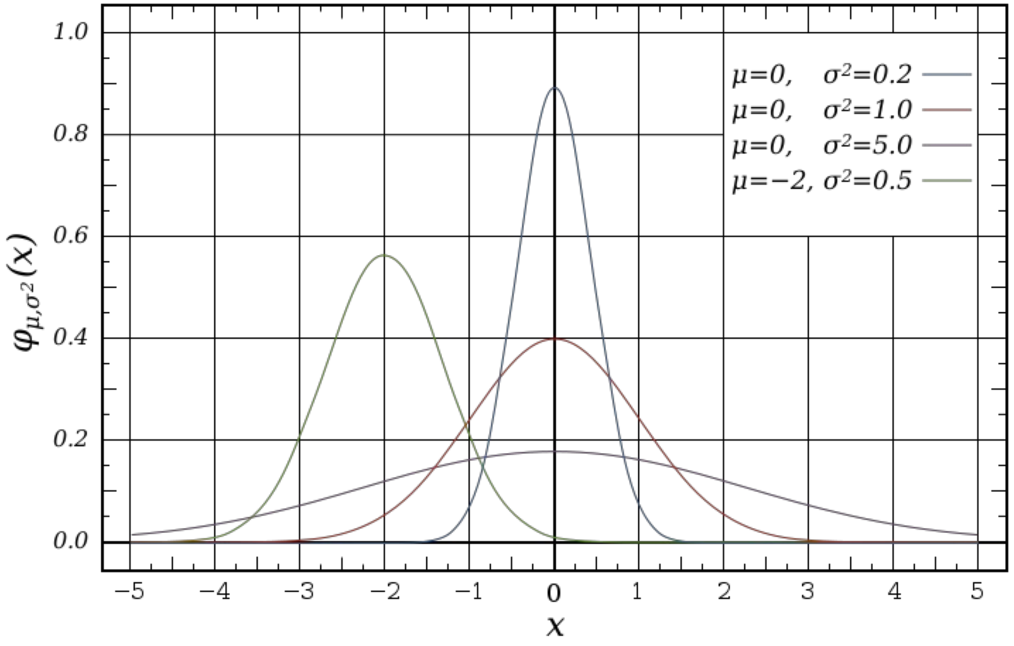
Some common distributions - Gaussian distribution

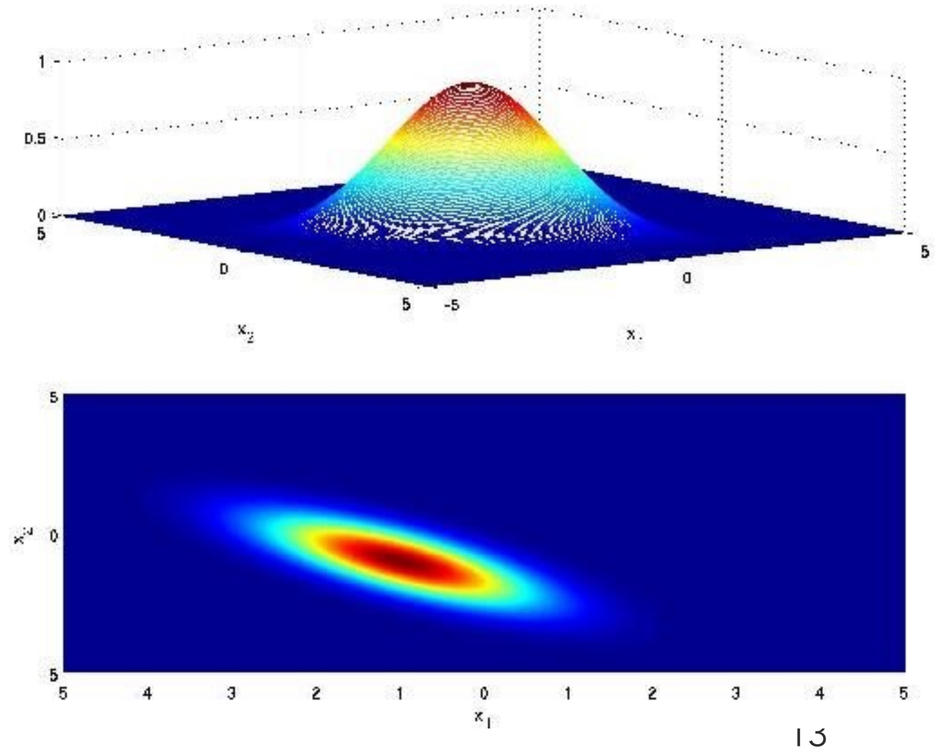
- Gaussian distribution

$$P(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

(Multivariate)

$$P(x \mid \mu, \Sigma) = \left|2\pi \Sigma\right|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$





[CSC2515, Wang]

Some common distributions - Multinomial distribution

- Multinomial distribution
  - Discrete random variable $\boldsymbol{x}$ that takes one of M values $\{1, \dots, M\}$
  - $p(\boldsymbol{x} = i) \;=\; \pi_i, \qquad \sum_i \pi_i = 1$
  - Out of n independent trials, let k_i be the number of times $\boldsymbol{x} = i$ was observed
  - The probability of observing a vector of occurrences $\boldsymbol{k} = [k_1, \dots, k_M]$ is given by the *multinomial distribution* parametrized by $\boldsymbol{\pi}$

$$p(\mathbf{k}|\boldsymbol{\pi}, n) \;=\; p(k_1, \dots, k_m|\pi_1, \dots, \pi_m, n) \;=\; \frac{n!}{k_1!k_2!\dots k_m!} \prod_{i=1} \pi_i^{k_i}$$

  - E.g., describing a text document by the frequency of occurrence of every distinct word
  - For $n = 1$, a.k.a. categorical distribution
    - $p(\boldsymbol{x} = i \mid \boldsymbol{\pi}) \;=\; \pi_i$
    - In $\boldsymbol{k} = [k_1, \dots, k_M]$: $k_i = 1$, and $k_j = 0$ for all $j \neq i \quad \rightarrow \quad a.k.a.$, one-hot representation of i

Exponential family

- A distribution

$$p_\theta(\boldsymbol{x}) = h(\boldsymbol{x}) \exp\{\boldsymbol{\theta} \cdot T(\boldsymbol{x})\} / Z(\boldsymbol{\theta})$$

  is an exponential family distribution

  - $\boldsymbol{\theta} \in R^d$: natural (canonical) parameter
  - $T(\boldsymbol{x}) \in R^d$: sufficient statistics, features of data $\boldsymbol{x}$
  - $Z(\boldsymbol{\theta}) = \sum_{x,y} h(\boldsymbol{x}) \exp\{\boldsymbol{\theta} \cdot T(\boldsymbol{x})\}$: normalization factor

- Examples: Bernoulli, multinomial, Gaussian, Poisson, gamma,...

Example: Multivariate Gaussian Distribution

- For a continuous vector random variable $\boldsymbol{x} \in R^k$

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

Moment parameter

$$= \frac{1}{(2\pi)^{k/2}} \exp\left\{-\tfrac{1}{2}\operatorname{tr}\left(\Sigma^{-1}xx^T\right) + \mu^T\Sigma^{-1}x - \tfrac{1}{2}\mu^T\Sigma^{-1}\mu - \log|\Sigma|\right\}$$

- Exponential family representation

$$\boldsymbol{\theta} = \left[\Sigma^{-1}\mu; -\frac{1}{2}\operatorname{vec}\left(\Sigma^{-1}\right)\right] = \left[\boldsymbol{\theta}_1, \operatorname{vec}\left(\boldsymbol{\theta}_2\right)\right], \ \boldsymbol{\theta}_1 = \Sigma^{-1}\mu \text{ and } \boldsymbol{\theta}_2^{-} = -\frac{1}{2}\Sigma^{-1}$$

$$T(\boldsymbol{x}) = \left[\boldsymbol{x}; \operatorname{vec}\left(\boldsymbol{x}\boldsymbol{x}^T\right)\right]$$

$$A(\boldsymbol{\theta}) = \frac{1}{2}\mu^T\Sigma^{-1}\mu + \log|\Sigma| = -\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\theta}_2\boldsymbol{\theta}_1\boldsymbol{\theta}_1^T\right) - \frac{1}{2}\log\left(-2\boldsymbol{\theta}_2\right)$$

$$h(x) = (2\pi)^{-k/2}$$

Probabilistic Graphical Models

Example

- Consider three binary-valued random variables

$$X_1, X_2, X_3 \qquad \mathrm{Val}(X_i) = \{0, 1\}$$

- Let outcome space Ω be the cross-product of their states:

$$\Omega = \mathrm{Val}(X_1) \times \mathrm{Val}(X_2) \times \mathrm{Val}(X_3)$$

- $X_i(\omega)$ is the value for X_i in the assignment $\omega \in \Omega$
- Specify $p(\omega)$ for each outcome $\omega \in \Omega$ by a big table:

x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	.11
0	0	1	.02
	$\vdots$		
1	1	1	.05

- How many parameters do we need to specify?

$$2^3 - 1$$

Marginalization

- Suppose X and Y are random variables with distribution $p(X, Y)$
  X: Intelligence, $\quad \text{Val}(X) = \{\text{"Very High"}, \text{"High"}\}$
  Y: Grade, $\quad \text{Val}(Y) = \{\text{"a"}, \text{"b"}\}$
- Joint distribution specified by:

X

Y		vh	h
	a	0.7	0.15
	b	0.1	0.05

- $p(Y = a) = ? = 0.85$
- More generally, suppose we have a joint distribution $p(X_1, \ldots, X_n)$. Then,

$$p(X_i = x_i) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} p(x_1, \ldots, x_n)$$

Conditioning

- Suppose X and Y are random variables with distribution $p(X, Y)$
  X: Intelligence, $\quad \text{Val}(X) = \{\text{"Very High"}, \text{"High"}\}$
  Y: Grade, $\quad \text{Val}(Y) = \{\text{"a"}, \text{"b"}\}$

Y \ X	vh	h
a	0.7	0.15
b	0.1	0.05

- Can compute the conditional probability

$$
\begin{aligned}
p(Y = a \mid X = vh) &= \frac{p(Y = a, X = vh)}{p(X = vh)} \\
&= \frac{p(Y = a, X = vh)}{p(Y = a, X = vh) + p(Y = b, X = vh)} \\
&= \frac{0.7}{0.7 + 0.1} \quad = 0.875.
\end{aligned}
$$

[PGM 2013, NYU, Sontag]

Example: Medical diagnosis

- Variable for each **symptom** (e.g. "fever", "cough", "fast breathing", "shaking", "nausea", "vomiting")
- Variable for each **disease** (e.g. "pneumonia", "flu", "common cold", "bronchitis", "tuberculosis")
- Diagnosis is performed by **inference** in the model:

$$p(\text{pneumonia} = 1 \mid \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0)$$

- One famous model, Quick Medical Reference (QMR-DT), has 600 diseases and 4000 findings

[PGM 2013, NYU, Sontag]

Representing the distribution

- Naively, could represent multivariate distributions with table of probabilities for each outcome (assignment)
- How many outcomes are there in QMR-DT? 2^{4600}
- **Estimation** of joint distribution would require a huge amount of data
- **Inference** of conditional probabilities, e.g.

$$p(\mathrm{pneumonia} = 1 \mid \mathrm{cough} = 1, \mathrm{fever} = 1, \mathrm{vomiting} = 0)$$

  would require summing over exponentially many variables' values
- Moreover, defeats the purpose of probabilistic modeling, which is to make predictions with *previously unseen observations*

Structure through independence

- If $X_1, \ldots, X_n$ are independent, then

$$p(x_1, \ldots, x_n) = p(x_1)p(x_2)\cdots p(x_n)$$

- 2^n entries can be described by just n numbers (if $|\mathrm{Val}(X_i)| = 2$)!
- However, this is not a very *useful* model – observing a variable X_i cannot influence our predictions of X_j
- If $X_1, \ldots, X_n$ are *conditionally independent* given Y, denoted as $X_i \perp \mathbf{X}_{-i} \mid Y$, then

$$\begin{aligned} p(y, x_1, \ldots, x_n) &= p(y)p(x_1 \mid y)\prod_{i=2}^{n} p(x_i \mid x_1, \ldots, x_{i-1}, y) \\ &= p(y)p(x_1 \mid y)\prod_{i=2}^{n} p(x_i \mid y). \end{aligned}$$

Bayesian networks (directed PGMs)

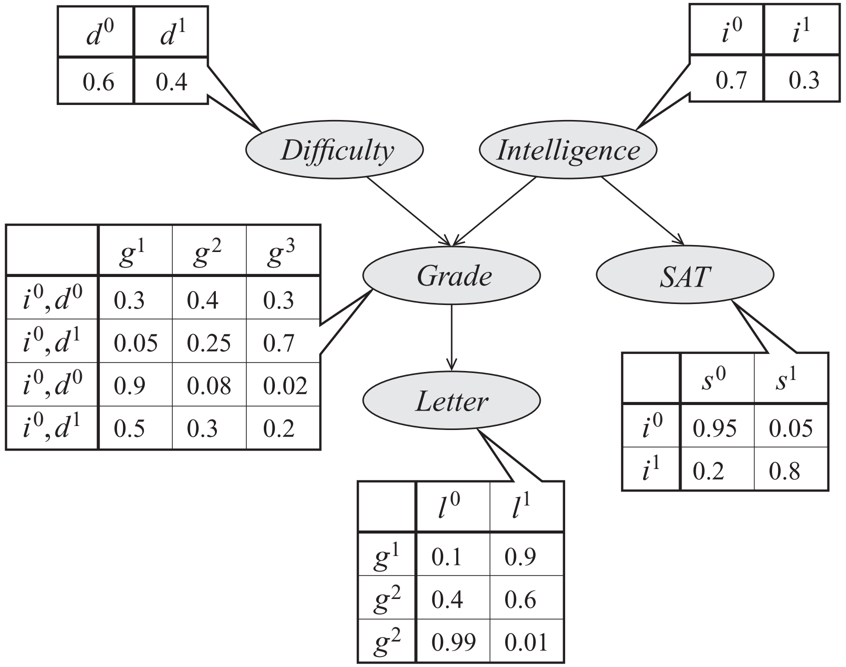
- A **Bayesian network** is specified by a directed *acyclic* graph $G = (V, E)$ with:
  1. One node $i \in V$ for each random variable X_i
  2. One conditional probability distribution (CPD) per node, $p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$, specifying the variable's probability conditioned on its parents' values
- Corresponds 1-1 with a particular factorization of the joint distribution:

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

- Powerful framework for designing *algorithms* to perform probability computations

Example

- Consider the following Bayesian network:



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	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^0,d^0	0.9	0.08	0.02
i^0,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^2	0.99	0.01

- What is its joint distribution?

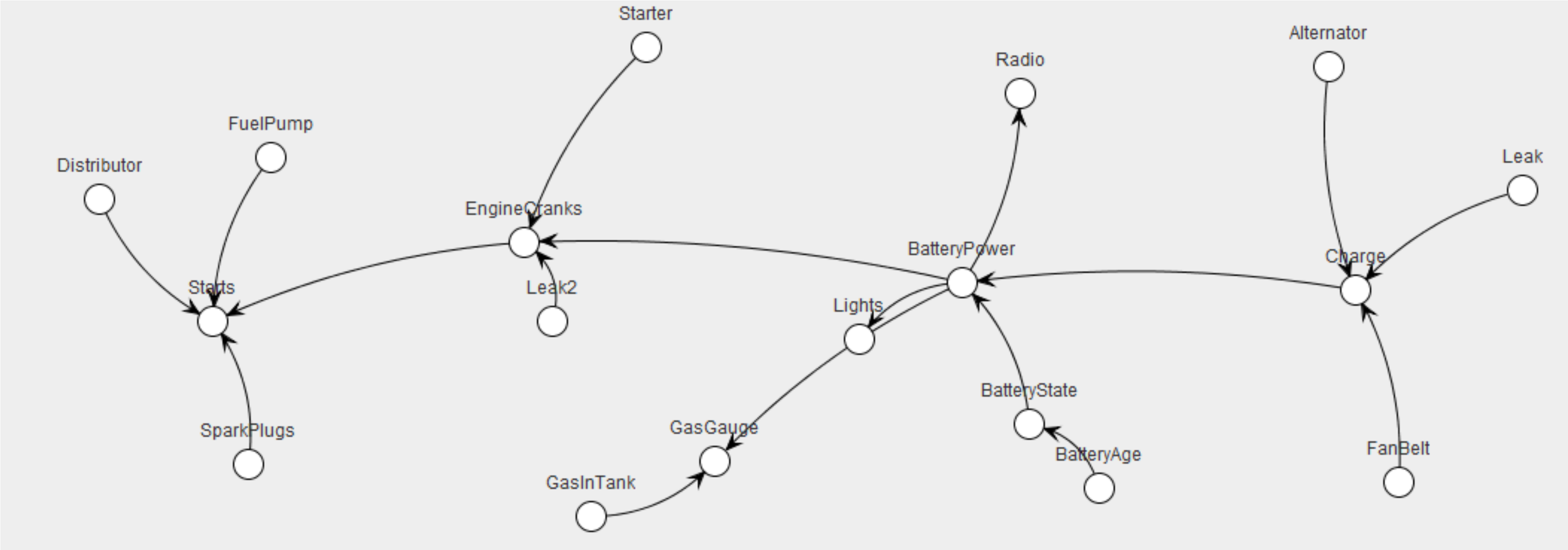
$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

$$p(d, i, g, s, l) = p(d)p(i)p(g \mid i, d)p(s \mid i)p(l \mid g)$$

More Examples

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

Will my car start this morning?



[Sontag] Heckerman *et al.*, Decision-Theoretic Troubleshooting, 1995

More Examples

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

What is the differential diagnosis?

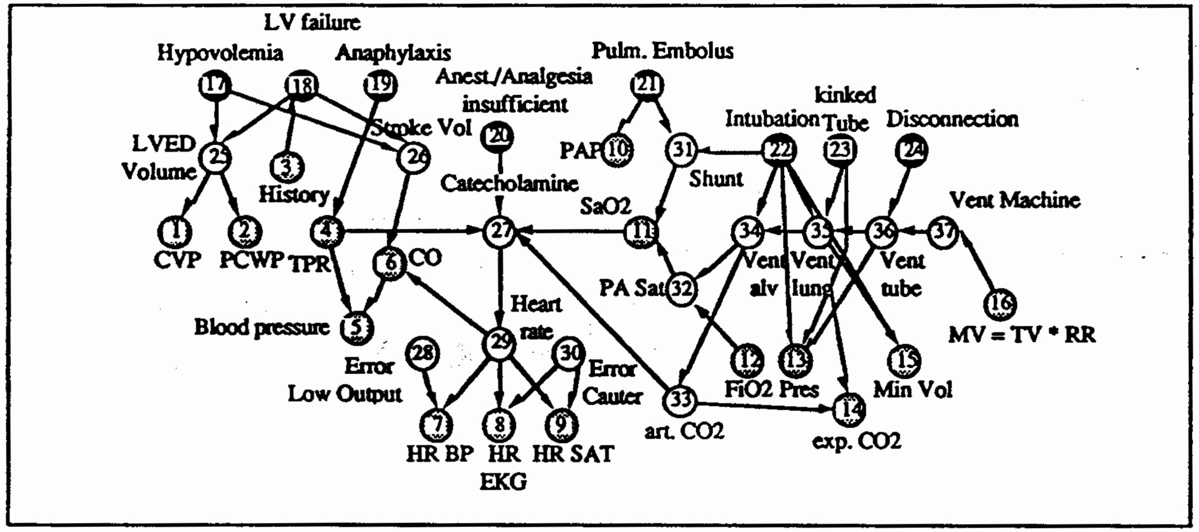


Fig. 1 The ALARM network representing causal relationships is shown with diagnostic (●), intermediate (○) and measurement (⊙) nodes. CO: cardiac output, CVP: central venous pressure, LVED volume: left ventricular end-diastolic volume, LV failure: left ventricular failure, MV: minute ventilation, PA Sat: pulmonary artery oxygen saturation, PAP: pulmonary artery pressure, PCWP: pulmonary capillary wedge pressure, Pres: breathing pressure, RR: respiratory rate, TPR: total peripheral resistance, TV: tidal volume

[Sontag] Beinlich *et al.*, The ALARM Monitoring System, 1989

Entropy

Entropy

- Shannon entropy $H(p) = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x})$
  - The average level of "information", "surprise", or "uncertainty" inherent to the variable $\boldsymbol{x}$'s possible outcomes

KL Divergence

- Kullback-Leibler (KL) divergence: measures the closeness of two distributions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$

$$\mathrm{KL}(q(\boldsymbol{x}) \,||\, p(\boldsymbol{x})) = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}$$

  - a.k.a. Relative entropy
  - KL >= 0 (Jensen's inequality)
  - Intuitively:
    - If q is high and p is high, then we are happy (i.e. low KL divergence)
    - If q is high and p is low then we pay a price (i.e. high KL divergence).
    - If q is low then we don't care (i.e. also low KL divergence, regardless of p)
  - not a true "distance":
    - not commutative (symmetric) $\mathrm{KL(p||q)} \mathrel{!}= \mathrm{KL(q||p)}$
    - doesn't satisfy triangle inequality

KL Divergence

- Kullback-Leibler (KL) divergence: measures the closeness of two distributions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$

$$\mathrm{KL}(q(\boldsymbol{x}) \,||\, p(\boldsymbol{x})) = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}$$

  - a.k.a. Relative entropy

- Maximum likelihood estimation (MLE) is minimizing the KL divergence between the empirical data distribution and the model distribution

$$\mathrm{KL}(\tilde{p}(\boldsymbol{x}) \,||\, p_\theta(\boldsymbol{x})) = \underbrace{-\mathbb{E}_{\tilde{p}(\boldsymbol{x})}[\log p_\theta(\boldsymbol{x})]}_{\text{Cross entropy}} + H(\tilde{p}(\boldsymbol{x}))$$

Key Takeaways

- Probability $p(\boldsymbol{x})$
- Bayes' rule $$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})}{p(\boldsymbol{x})}$$
  - prior, posterior
- Exponential family:
  - Gaussian, multinomial, categorical, …
- Probabilistic graphical models: Bayesian networks
- KL Divergence
  - relation to Cross-entropy

$$\mathrm{KL}(q(\boldsymbol{x}) \,||\, p(\boldsymbol{x})) = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}$$

Functional Derivatives

Functional derivative

- $\nabla_q - \mathbb{H}(q) = \log q + 1$

- Functional $F(y)$: an operator that takes a function $y(x)$ and returns an output value F
- Functional derivative (aka, variational derivative): relates a change in a Functional $F(y)$ to a change in the function y

Functional derivative

- Recall the conventional derivative $\frac{dy}{dx}$
  - Taylor expansion

$$y(x+\epsilon) = y(x) + \frac{\mathrm{d}y}{\mathrm{d}x}\epsilon + O(\epsilon^2)$$

- Functional derivative
  - How much a functional $F[y]$ changes when we make a small change $\varepsilon\eta(x)$ to the function $y(x)$

$$F[y(x) + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\delta F}{\delta y(x)} \eta(x)\, \mathrm{d}x + O(\epsilon^2)$$

  - A function $y(x)$ that maximizes (or minimizes) a functional $F[y]$ must satisfy

$$\frac{\delta F}{\delta y(x)} = 0 \text{ for all } x$$

Functional derivative

$$F[y(x) + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\delta F}{\delta y(x)} \eta(x)\,\mathrm{d}x + O(\epsilon^2)$$

- Consider a functional that is defined by an integral over a function $G(y, x)$

$$F[y] = \int G(y, x) dx$$

- Consider variations in the function $y(x)$,

$$F[y + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\partial G}{\partial y} \eta(x) dx + O(\epsilon^2)$$

Functional derivative

$$F[y(x) + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\delta F}{\delta y(x)} \eta(x)\, \mathrm{d}x + O(\epsilon^2)$$

- Consider a functional that is defined by an integral over a function $G(y, x)$
$$F[y] = \int G(y, x) dx$$
  - Ex.1, $-\mathbb{H}(q) = \int q(x) \log q(x)\, dx$
    - $G = q(x) \log q(x)$
- Consider variations in the function $y(x)$,

$$F[y + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\partial G}{\partial y} \eta(x) dx + O(\epsilon^2)$$

Practice: Maximum likelihood vs Maximum Entropy

Supervised Maximum Likelihood

- Model to be learned $p_\theta(\boldsymbol{x})$
- Observe full data $\mathcal{D} = \{ \boldsymbol{x}^* \}$
  - i.i.d: independent, identically distributed
- Maximum Likelihood Estimation (MLE)
  - The most classical learning algorithm

$$\min_\theta - \mathbb{E}_{x^* \sim \mathcal{D}} \left[\quad \log p_\theta(\boldsymbol{x}^*) \quad \right]$$

- MLE is closely connected to the Maximum Entropy (MaxEnt) principle

Recap: Exponential Family

- A distribution

$$p_\theta(\boldsymbol{x}) = h(\boldsymbol{x}) \exp\{\boldsymbol{\theta} \cdot T(\boldsymbol{x})\} / Z(\boldsymbol{\theta})$$

is an exponential family distribution

  - $\boldsymbol{\theta} \in R^d$: natural (canonical) parameter
  - $T(\boldsymbol{x}) \in R^d$: sufficient statistics, features of data $\boldsymbol{x}$
  - $Z(\boldsymbol{\theta}) = \sum_{x,y} h(\boldsymbol{x}) \exp\{\boldsymbol{\theta} \cdot T(\boldsymbol{x})\}$: normalization factor

- Examples: Bernoulli, multinomial, Gaussian, Poisson, gamma,...

Maximum Likelihood for Exponential Family

$m(\boldsymbol{x})$: the number of times $\boldsymbol{x}$ is observed in *D*

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta};\mathcal{D}) &= \sum_{\boldsymbol{x}} m(\boldsymbol{x}) \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) \\ &= \sum_{\boldsymbol{x}} m(\boldsymbol{x}) \left(\sum_i \theta_i T_i(\boldsymbol{x}) - \log Z(\boldsymbol{\theta}) \right) \\ &= \sum_{\boldsymbol{x}} m(\boldsymbol{x}) \sum_i \theta_i T_i(\boldsymbol{x}) - N \log Z(\boldsymbol{\theta})\end{aligned}$$

- Take gradient and set to 0

$$\begin{aligned}\frac{\partial}{\partial \theta_i}\mathcal{L}(\boldsymbol{\theta};\mathcal{D}) &= \sum_{\boldsymbol{x}} m(\boldsymbol{x}) T_i(\boldsymbol{x}) - N \frac{\partial}{\partial \theta_i} \log Z(\boldsymbol{\theta}) \\ &= \sum_{\boldsymbol{x}} m(\boldsymbol{x}) T_i(\boldsymbol{x}) - N \sum_{\boldsymbol{x}} p(\boldsymbol{x} \mid \boldsymbol{\theta}) T_i(\boldsymbol{x})\end{aligned}$$

$$\Rightarrow \boxed{\sum_{\boldsymbol{x}} p(\boldsymbol{x} \mid \boldsymbol{\theta}) T_i(\boldsymbol{x})} = \sum_{\boldsymbol{x}} \frac{m(\boldsymbol{x})}{N} T_i(\boldsymbol{x}) = \boxed{\sum_{\boldsymbol{x}} \tilde{p}(\boldsymbol{x} \mid \boldsymbol{\theta}) T_i(\boldsymbol{x})}$$

At MLE, the expectations of the sufficient statistics under the model must match empirical feature average

Maximum Entropy (MaxEnt)

- Given $\mathcal{D}$, to estimate $p(\boldsymbol{x})$
- We can approach the problem from an entirely different point of view. Begin with some fixed feature expectations:

$$\sum_{\boldsymbol{x}} p(\boldsymbol{x}) T_i(\boldsymbol{x}) = \sum_{\boldsymbol{x}} \frac{m(\boldsymbol{x})}{N} T_i(\boldsymbol{x}) \coloneqq \alpha_i$$

- There may exist many distributions which satisfy them. Which one should we select?
  - MaxEnt principle: the most uncertain or flexible one, i.e., the one with maximum entropy
- This yields a new optimization problem:
  - This is a <u>variational</u> definition of a distribution!

$$\boxed{\begin{aligned} \max_p \;\; & \mathrm{H}(p(\boldsymbol{x})) = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) \\ \text{s.t.} \;\; & \sum_{\boldsymbol{x}} p(\boldsymbol{x}) T_i(\boldsymbol{x}) = \alpha_i \\ & \sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1 \end{aligned}}$$

Solution to the MaxEnt Problem

- To solve the MaxEnt problem, we use Lagrange multipliers:

$$\max_{\theta,\mu} \min_{p(x)} \; L = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) - \sum_{i} \theta_i \left(\sum_{x} p(\boldsymbol{x}) T_i(\boldsymbol{x}) - \alpha_i \right) - \mu \left(\sum_{\boldsymbol{x}} p(\boldsymbol{x}) - 1 \right)$$

Solution to the MaxEnt Problem

- To solve the MaxEnt problem, we use Lagrange multipliers:

$$\max_{\theta,\mu} \min_{p(x)} \; L = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) - \sum_i \theta_i \left(\sum_x p(\boldsymbol{x}) T_i(\boldsymbol{x}) - \alpha_i \right) - \mu \left(\sum_{\boldsymbol{x}} p(\boldsymbol{x}) - 1 \right)$$

$$\frac{\partial L}{\partial p(\boldsymbol{x})} = 1 + \log p(\boldsymbol{x}) - \sum_i \theta_i T_i(\boldsymbol{x}) - \mu$$

$$p^*(\boldsymbol{x}) = e^{\mu - 1} \exp\left\{ \sum_i \theta_i f_i(\boldsymbol{x}) \right\}$$

$$Z(\boldsymbol{\theta}) = e^{\mu - 1} = \sum_{\boldsymbol{x}} \exp\left\{ \sum_i \theta_i f_i(\boldsymbol{x}) \right\} \quad \left(\text{since } \sum_{\boldsymbol{x}} p^*(\boldsymbol{x}) = 1 \right)$$

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left\{ \sum_i \theta_i T_i(\boldsymbol{x}) \right\}$$

Solution to the MaxEnt Problem

- To solve the MaxEnt problem, we use Lagrange multipliers:

$$\max_{\theta,\mu} \min_{p(x)} L = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) - \sum_i \theta_i \left(\sum_x p(\boldsymbol{x}) T_i(\boldsymbol{x}) - \alpha_i \right) - \mu \left(\sum_{\boldsymbol{x}} p(\boldsymbol{x}) - 1 \right)$$

$$\frac{\partial L}{\partial p(\boldsymbol{x})} = 1 + \log p(\boldsymbol{x}) - \sum_i \theta_i T_i(\boldsymbol{x}) - \mu$$

$$p^*(\boldsymbol{x}) = e^{\mu-1} \exp\left\{ \sum_i \theta_i f_i(\boldsymbol{x}) \right\}$$

$$Z(\boldsymbol{\theta}) = e^{\mu-1} = \sum_{\boldsymbol{x}} \exp\left\{ \sum_i \theta_i f_i(\boldsymbol{x}) \right\} \quad \left(\text{since} \sum_{\boldsymbol{x}} p^*(\boldsymbol{x}) = 1 \right)$$

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left\{ \sum_i \theta_i T_i(\boldsymbol{x}) \right\}$$

- So feature constraints + MaxEnt $\Rightarrow$ exponential family.
- Problem is strictly convex w.r.t. $p(\boldsymbol{x})$, so solution is unique.

Solution to the MaxEnt Problem

- To solve the MaxEnt problem, we use Lagrange multipliers:

$$\max_{\theta,\mu} \min_{p(x)} \; L = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) - \sum_i \theta_i \left(\sum_x p(\boldsymbol{x}) T_i(\boldsymbol{x}) - \alpha_i \right) - \mu \left(\sum_{\boldsymbol{x}} p(\boldsymbol{x}) - 1 \right)$$

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp \left\{ \sum_i \theta_i T_i(\boldsymbol{x}) \right\}$$

- So feature constraints + MaxEnt $\Rightarrow$ exponential family.
- Problem is strictly convex w.r.t. $p(\boldsymbol{x})$, so solution is unique.

plug $p(x|\boldsymbol{\theta})$ back into L, and since $\sum_{\boldsymbol{x}} \frac{m(\boldsymbol{x})}{N} T_i(\boldsymbol{x}) \coloneqq \alpha_i$:

$$\max_{\theta} L(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} m(\boldsymbol{x}) \sum_i \theta_i T_i(\boldsymbol{x}) - N \log Z(\boldsymbol{\theta})$$

- Recovers precisely the MLE problem of exponential family

(Homework)

Constraints from Data

- We have seen a case of **convex duality**:
  - In one case, we assume exponential family and show that Maximum Likelihood implies model expectations must match empirical expectations.
  - In the other case, we assume model expectations must match empirical feature counts and show that MaxEnt implies exponential family distribution.

A more general MaxEnt problem

$$\min_p \quad \mathrm{KL}(p(\boldsymbol{x}) \| h(\boldsymbol{x}))$$

$$\stackrel{\text{def}}{=} \sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{h(\boldsymbol{x})} = -\mathrm{H}(p) - \sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log h(\boldsymbol{x})$$

$$\text{s.t.} \quad \sum_{\boldsymbol{x}} p(\boldsymbol{x}) T_i(\boldsymbol{x}) = \alpha_i$$

$$\sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1$$

$$\Rightarrow \quad p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} h(\boldsymbol{x}) \exp\left\{ \sum_i \theta_i T_i(\boldsymbol{x}) \right\}$$

Summary

- Maximum entropy is dual to maximum likelihood of exponential family distributions
- This provides an alternative view of the problem of fitting a model into data:
  - The data instances in the training set are treated as constraints, and the learning problem is treated as a constrained optimization problem.
  - We'll revisit this optimization-theoretic view of learning repeatedly in the future!

$$\boxed{\begin{aligned} \max_{p} \quad & \mathrm{H}(p(\boldsymbol{x})) = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) \\ \text{s.t.} \quad & \sum_{\boldsymbol{x}} p(\boldsymbol{x}) T_i(\boldsymbol{x}) = \alpha_i \\ & \sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1 \end{aligned}}$$

Key Takeaways

- Probability
    - Bayes' rule
    - Exponential family
    - Probabilistic graphical models: Bayesian networks
    - KL divergence
- Functional derivative
- Convex duality between MLE and MaxEnt (optional)

Questions?