15-780 – Graduate Artificial Intelligence: Probabilistic modeling

J. Zico Kolter (this lecture) and Ariel Procaccia
Carnegie Mellon University
Spring 2017

Outline

Probability in AI

Background on probability

Common distributions

Maximum likelihood estimation

Probabilistic graphical models

Outline

Probability in AI

Background on probability

Common distributions

Maximum likelihood estimation

Probabilistic graphical models

Probability in AI

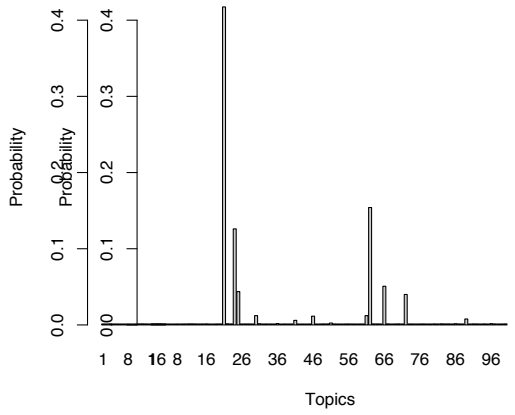
Basic idea: the real world is probabilistic (at least at the level we can observe it), and our reasoning about it needs to be too

The shift from “logical” to “probabilistic” AI systems (circa 80s, 90s) represented a revolution in AI

Probabilistic approaches are now intertwined with virtually all areas of AI

Example: topic modeling

Can we learn about the content of text documents just be reading through them and see what sorts of words "co-occur"



"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

Figure from (Blei et al., 2011) demonstrates words and topics recovered from reading 17,000 *Science* articles

Example: biological networks

Can we automatically determine how the presence or absence of some proteins in a cell affect others?

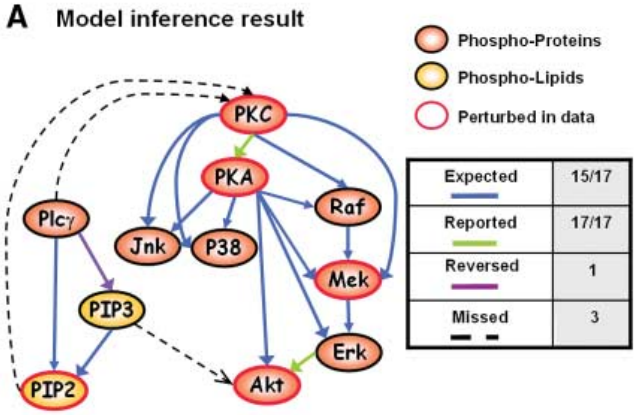


Figure from (Sachs et al., 2005) shows automatically inferred protein probability network, which captured most of the known interactions using data-driven methods (far less manual effort than previous methods)

Outline

Probability in AI

Background on probability

Common distributions

Maximum likelihood estimation

Probabilistic graphical models

Random variables

A random variable (informally) is a variable whose value is not initial known

Instead, these variables can take on different values (including a possibly infinite number), and must take on exactly one of these values, each with an associated probability, which all together sum to one

$$\text{“Weather” takes values } \{\text{sunny}, \text{rainy}, \text{cloudy}, \text{snowy}\}$$
$$p(\text{Weather} = \text{ sunny}) = 0.3$$
$$p(\text{Weather} = \text{rainy}) = 0.2$$
$$\ldots$$

Slightly different notation for continuous random variables, which we will discuss shortly

Notation for random variables

In this lecture, we use upper case letters, X_i to denote random variables

For a random variable X_i taking values $\{1,2,3\}$

$$p(X_i) = \begin{pmatrix} 0.1 \\ 0.5 \\ 0.4 \end{pmatrix}$$

represents the set of probabilities for each value that X_i can take on (this is a *function* mapping values of X_i to numbers that sum to one)

Conversely, we will use lower case x_i to denote a specific *value* of X_i (i.e., for above example $x_i \in \{1,2,3\}$), and $p(X_i = x_i)$ or just $p(x_i)$ refers to a *number* (the corresponding entry of $p(X_i)$)

Examples of probability notation

Given two random variables: X_1 with values in $\{1,2,3\}$ and X_2 with values in $\{1,2\}$:

$p(X_1, X_2)$ refers to the *joint distribution,* i.e., a set of 6 possible values for each setting of variables, i.e. a function mapping $(1,1), (1,2), (2,1), \ldots$ to corresponding probabilities)

$p(x_1, x_2)$ is a *number:* probability that $X_1 = x_1$ and $X_2 = x_2$

$p(X_1, x_2)$ is a set of 3 values, the probabilities for all values of X_1 for the given value $X_2 = x_2$, i.e., it is a function mapping $0,1,2$ to numbers (note: *not* probability distribution, it will not sum to one)

We generally call all of these terms ***factors*** (functions mapping values to numbers, even if they do not sum to one)

Operations on probabilities/factors

We can perform operations on probabilities/factors by performing the operation on every corresponding value in the probabilities/factors

For example, given three random variables X_1, X_2, X_3:

$$p(X_1, X_2) \ \langle \text{op} \rangle \ p(X_2, X_3)$$

denotes a factor over X_1, X_2, X_3 (i.e., a function over all possible combinations of values these three random variables can take), where the value for x_1, x_2, x_3 is given by

$$p(x_1, x_2) \ \langle \text{op} \rangle \ p(x_2, x_3)$$

Conditional probability

The ***conditional probability*** $p(X_1|X_2)$ (the conditional probability of X_1 given X_2) is defined as

$$p(X_1|X_2) = \frac{p(X_1, X_2)}{p(X_2)}$$

Can also be written $p(X_1, X_2) = p(X_1|X_2)p(X_2)$

More generally, leads to the ***chain rule:***

$$p(X_1, \dots, X_n) = \prod_{i=1}^{n} p(X_i|X_1, \dots X_{i-1})$$

Marginalization

For random variables X_1, X_2 with joint distribution $p(X_1, X_2)$

$$p(X_1) = \sum_{x_2} p(X_1, x_2) = \sum_{x_2} p(X_1|x_2)p(x_2)$$

Generalizes to joint distributions over multiple random variables

$$p(X_1, \dots, X_i) = \sum_{x_{i+1}, \dots, x_n} p(X_1, \dots, X_i, x_{i+1}, \dots, x_n)$$

For p to be a probability distribution, the marginalization over *all* variables must be one

$$\sum_{x_1, \dots, x_n} p(x_1, \dots, x_n) = 1$$

Bayes' rule

A straightforward manipulation of probabilities:

$$p(X_1|X_2) = \frac{p(X_1, X_2)}{p(X_2)} = \frac{p(X_2|X_1)p(X_1)}{p(X_2)} = \frac{p(X_2|X_1)p(X_1)}{\sum_{x_1} p(X_2|x_1)\, p(x_1)}$$

Poll: I want to know if I have come down with a rate strain of flu (occurring in only 1/10,000 people). There is an "accurate" test for the flu: if I have the flu, it will tell me I have 99% of the time, and if I do not have it, it will tell me I do not have it 99% of the time. I go to the doctor and test positive. What is the probability I have the this flu?

1. $\approx 99\%$
2. $\approx 10\%$
3. $\approx 1\%$
4. $\approx 0.1\%$

Independence

We say that random variables X_1 and X_2 are ***(marginally) independent*** if their joint distribution is the product of their marginals

$$p(X_1, X_2) = p(X_1)p(X_2)$$

Equivalently, can also be stated as the condition that

$$p(X_1|X_2) \quad \left(= \frac{p(X_1, X_2)}{p(X_2)} = \frac{p(X_1)p(X_2)}{p(X_2)} \right) = p(X_1)$$

$$(\text{and similarly}) \quad p(X_2|X_1) = p(X_2)$$

Conditional independence

We say that random variables X_1 and X_2 are ***conditionally independent given*** X_3, if

$$p(X_1, X_2|X_3) = p(X_1|X_3)p(X_2|X_3)$$

Again, can be equivalently written:

$$p(X_1|X_2, X_3) \left(= \frac{p(X_1, X_2|X_3)}{p(X_2|X_3)} = \frac{p(X_1|X_3)p(X_2|X_3)}{p(X_2|X_3)} \right)$$
$$= p(X_1|X_3)$$

And similarly $p(X_2|X_1, X_3) = p(X_2|X_3)$

Important: Marginal independence does not imply conditional independence or vice versa

Expectation

The expectation of a random variable is denoted:

$$\mathbf{E}[X] = \sum_x x \cdot p(x)$$

where we use upper case X to emphasize that this is a function of the entire random variable (but unlike $p(X)$ is a number)

Note that this only makes sense when the values that the random variable takes on are *numerical* (i.e., We can't ask for the expectation of the random variable "Weather")

Also generalizes to *conditional expectation:*

$$\mathbf{E}[X_1|x_2] = \sum_{x_1} x_1 \cdot p(x_1|x_2)$$

Rules of expectation

Expectation of sum is always equal to sum of expectations (even when variables are not independent):

$$\begin{aligned}\mathbf{E}[X_1 + X_2] &= \sum_{x_1,x_2} (x_1 + x_2) p(x_1, x_2) \\ &= \sum_{x_1} x_1 \sum_{x_2} p(x_1, x_2) + \sum_{x_2} x_2 \sum_{x_1} p(x_1, x_2) \\ &= \sum_{x_1} x_1 p(x_1) + \sum_{x_2} x_2 p(x_2) = \mathbf{E}[X_1] + \mathbf{E}[X_2]\end{aligned}$$

If x_1, x_2 independent, expectation of products is product of expectations

$$\begin{aligned}\mathbf{E}[X_1 X_2] &= \sum_{x_1,x_2} x_1 x_2 \, p(x_1, x_2) = \sum_{x_1,x_2} x_1 x_2 \, p(x_1) p(x_2) \\ &= \sum_{x_1} x_1 p(x_1) \sum_{x_2} x_2 p(x_2) = \mathbf{E}[X_1]\mathbf{E}[X_2]\end{aligned}$$

Variance

Variance of a random variable is the expectation of the variable minus its expectation, squared

$$\begin{aligned}\mathbf{Var}[X] &= \mathbf{E}[(X - \mathbf{E}[X])^2] \left(= \sum_x (x - \mathbf{E}[x])^2 p(x) \right) \\ &= \mathbf{E}[X^2 - 2X\mathbf{E}[X] + \mathbf{E}[X]^2] = \mathbf{E}[X^2] - \mathbf{E}[X]^2\end{aligned}$$

Generalizes to covariance between two random variables

$$\begin{aligned}\mathbf{Cov}[X_1, X_2] &= \mathbf{E}[(X_1 - \mathbf{E}[X_1])(X_2 - \mathbf{E}[X_2])] \\ &= \mathbf{E}[X_1 X_2] - \mathbf{E}[X_1]\mathbf{E}[X_2]\end{aligned}$$

Infinite random variables

All the math above works the same for discrete random variables that can take on an infinite number of values (I'm talking about *countably infinite* values here)

The only difference is that $p(X)$ (obviously) cannot be specified by an explicit dictionary mapping variable values to probabilities, need to specify the functional form that produces probabilities

To be a probability, we still must have $\sum_x p(x) = 1$

Example:

$$P(X = k) = \left(\frac{1}{2}\right)^k, \qquad k = 1, \ldots, \infty$$

Continuous random variables

For random variables taking on *continuous* values (we'll only consider real-valued distributions), we need some slightly different mechanisms

As with infinite discrete variables, the distribution $p(X)$ needs to be specified as a function: here is referred to as a ***probability density function*** (PDF) and it must *integrate* to one $\int_{\mathbb{R}} p(x)dx = 1$

For any interval (a, b), we have that $p(a \leq x \leq b) = \int_a^b p(x)dx$ (with similar generalization to multi-dimensional random variables)

Can also be specified by their ***cumulative distribution function*** (CDF),

$$F(a) = p(x \leq a) = \int_{\infty}^{a} p(x)$$

Outline

Probability in AI

Background on probability

Common distributions

Maximum likelihood estimation

Probabilistic graphical models

Bernoulli distribution

A simple distribution over binary $\{0,1\}$ random variables

$$p(X=1;\phi)=\phi, \qquad P(X=0;\phi)=1-\phi$$

where $\phi \in [0,1]$ is the parameter that governs the distribution

Expectation is just $\mathbf{E}[x]=\phi$ (but not very common to refer to it this way, since this would imply that the $\{0,1\}$ terms are actual real-valued numbers)

Categorical distribution

This is the discrete distribution we've mainly considered so far, a distribute over finite discrete elements with each probability specified

Written generically as:

$$p(X = i; \phi) = \phi_i$$

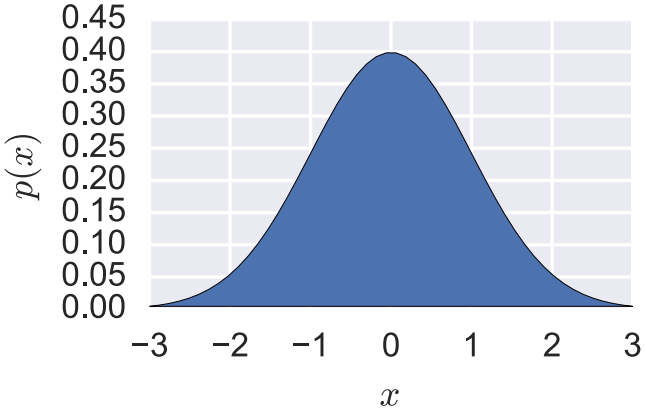
where $\phi_1, \ldots \phi_k \in [0,1]$ are the parameters of the distribution (the probability of each random variable, must sum to one)

Note: we could actually parameterize just using $\phi_1, \ldots \phi_{k-1}$, since this would determine the last elements

Unless the actual numerical value of the i's are relevant, it doesn't make sense to take expectations of a categorical random variable

Gaussian distribution

Distribution over real-valued numbers, empirically the most common distribution in all of data science (*not* in data itself, necessarily, but for people applying data science), the standard "bell curve":



$$\mu = 0$$
$$\sigma^2 = 1$$

Probability density function:

$$p(x; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \equiv \mathcal{N}(x; \mu, \sigma^2)$$

with parameters $\mu \in \mathbb{R}$ (mean) and $\sigma^2 \in \mathbb{R}_+$ (variance)

Multivariate Gaussians

The Gaussian distribution is one of the few distributions that generalizes nicely to higher dimensions

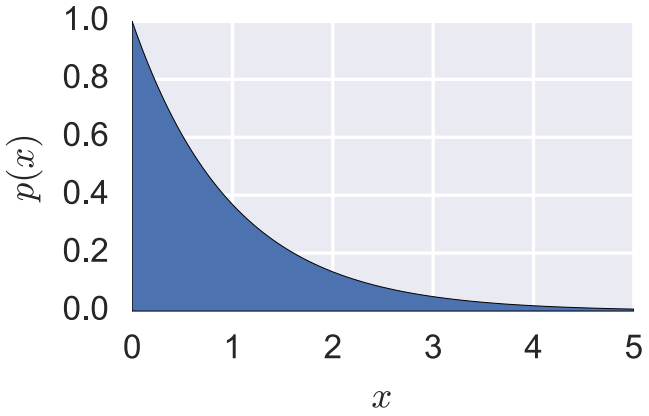
We'll discuss this in much more detail when we talk about anomaly detection and the mixture of Gaussians model, but for now, just know that we can also write a distribution over random *vectors* $x \in \mathbb{R}^n$

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

where $\mu \in \mathbb{R}^n$ is mean and $\Sigma \in \mathbb{R}^{n \times n}$ is *covariance matrix,* and $|\cdot|$ denotes the determinant of a matrix

Exponential distribution

A one-sided Laplace distribution, often used to model arrival times



$\lambda = 1$

Probability density function:

$$p(x; \lambda) = \lambda \exp(-\lambda x)$$

with parameter $\lambda \in \mathbb{R}_+$ (mean/variance $\mathbf{E}[X] = 1/\lambda, \mathbf{Var}[x] = 1/\lambda^2$)

Outline

Probability in AI

Background on probability

Common distributions

Maximum likelihood estimation

Probabilistic graphical models

Estimating the parameters of distributions

We're moving now from probability to statistics

The basic question: given some data $x^{(1)}, \ldots, x^{(m)}$, how do I find a distribution that captures this data "well"?

In general (if we can pick from the space of all distributions), this is a hard question, but if we pick from a particular *parameterized family* of distributions $p(X; \theta)$, the question is (at least a little bit) easier

Question becomes: how do I find parameters θ of this distribution that fit the data?

Maximum likelihood estimation

Given a distribution $p(X;\theta)$, and a collection of observed (independent) data points $x^{(1)}, \ldots, x^{(m)}$, the probability of observing this data is simply

$$p(x^{(1)}, \ldots, x^{(m)}; \theta) = \prod_{i=1}^{m} p(x^{(i)}; \theta)$$

Basic idea of maximum likelihood estimation (MLE): find the parameters that maximize the probability of the observed data

$$\underset{\theta}{\text{maximize}} \prod_{i=1}^{m} p(x^{(i)}; \theta) \equiv \underset{\theta}{\text{maximize}} \ \ell(\theta) = \sum_{i=1}^{m} \log p(x^{(i)}; \theta)$$

where $\ell(\theta)$ is called the ***log likelihood*** of the data

Seems “obvious”, but there are many other ways of fitting parameters

Parameter estimation for Bernoulli

Simple example: Bernoulli distribution

$$p(X = 1; \phi) = \phi, \qquad p(X = 0; \phi) = 1 - \phi$$

Given observed data $x^{(1)}, \ldots, x^{(m)}$, the "obvious" answer is:

$$\hat{\phi} = \frac{\#1\text{'s}}{\#\ \text{Total}} = \frac{\sum_{i=1}^{m} x^{(i)}}{m}$$

But why is this the case?

Maybe there are other estimates that are just as good, i.e.?

$$\phi = \frac{\sum_{i=1}^{m} x^{(i)} + 1}{m + 2}$$

MLE for Bernoulli

Maximum likelihood solution for Bernoulli given by

$$\underset{\phi}{\text{maximize}} \prod_{i=1}^{m} p(x^{(i)}; \phi) = \underset{\phi}{\text{maximize}} \prod_{i=1}^{m} \phi^{x^{(i)}} (1-\phi)^{1-x^{(i)}}$$

Taking the negative log of the optimization objective (just to be consistent with our usual notation of optimization as minimization)

$$\underset{\phi}{\text{maximize}}\ \ell(\phi) = \sum_{i=1}^{m} \left(x^{(i)} \log \phi + \left(1 - x^{(i)}\right) \log(1-\phi)\right)$$

Derivative with respect to ϕ is given by

$$\frac{d}{d\phi}\ell(\phi) = \sum_{i=1}^{m} \left(\frac{x^{(i)}}{\phi} - \frac{1-x^{(i)}}{1-\phi}\right) = \frac{\sum_{i=1}^{m} x^{(i)}}{\phi} - \frac{\sum_{i=1}^{m} (1-x^{(i)})}{1-\phi}$$

MLE for Bernoulli, continued

Setting derivative to zero gives:

$$\frac{\sum_{i=1}^{m} x^{(i)}}{\phi} - \frac{\sum_{i=1}^{m}(1 - x^{(i)})}{1-\phi} \equiv \frac{a}{\phi} - \frac{b}{1-\phi} = 0$$
$$\implies (1-\phi)a = \phi b$$
$$\implies \phi = \frac{a}{a+b} = \frac{\sum_{i=1}^{m} x^{(i)}}{m}$$

So, we have shown that the "natural" estimate of ϕ actually corresponds to the maximum likelihood estimate

MLE for Gaussian, briefly

For Gaussian distribution

$$p(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp(-(1/2)(x-\mu)^2/\sigma^2)$$

Log likelihood given by:

$$\ell(\mu, \sigma^2) = -m\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{m}\frac{(x^{(i)}-\mu)^2}{\sigma^2}$$

Derivatives (see if you can derive these fully):

$$\frac{d}{d\mu}\ell(\mu, \sigma^2) = -\frac{1}{2}\sum_{i=1}^{m}\frac{x^{(i)}-\mu}{\sigma^2} = 0 \Longrightarrow \mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\frac{d}{d\sigma^2}\ell(\mu, \sigma^2) = -\frac{m}{2\sigma^2} + \frac{1}{2}\sum_{i=1}^{m}\frac{(x^{(i)}-\mu)^2}{(\sigma^2)^2} = 0 \Longrightarrow \sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu)^2$$

Machine learning via maximum likelihood

Many machine learning algorithms (specifically the loss function component) can be interpreted as maximum likelihood estimation

Logistic regression:

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{m} \log(1 + \exp\left(-y^{(i)} \cdot h_\theta(x^{(i)})\right)$$

Softmax (multiclass logistic) regression:

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{m} \left(\log \sum_{j=1}^{k} \exp(h_\theta(x^{(i)})_j) - h_\theta(x^{(i)})^T y^{(i)} \right)$$

Where did these come from?

Logistic model

Suppose our random variable Y takes on values in $\{-1, +1\}$ and we want to model the condition distribution $p(Y|X)$ as a function of $\theta^T x$ for some parameters θ

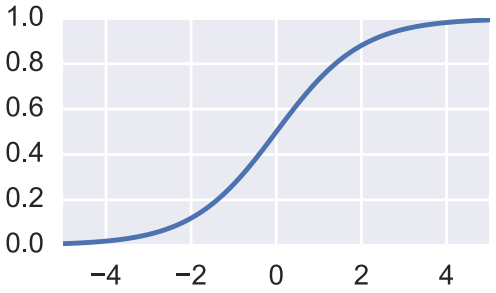
Since probabilities must be positive, let's look at the distribution

$$p(y = +1|x; \theta) \propto \exp(\theta^T x)\,, \qquad p(y = -1|x; \theta) \propto 1$$

Then, because the actual probability values need to sum to one

$$p(y = +1|x; \theta) = \frac{\exp(\theta^T x)}{1 + \exp(\theta^T x)} = \frac{1}{1 + \exp(-\theta^T x)}$$

This last term is called a *logistic* (or *sigmoid)* function $\sigma(z) = 1/(1 + \exp(-z))$



Logistic probability model

Under linear logistic model we can write the likelihood as

$$p(y|x;\theta) = \sigma(y \cdot \theta^T x)$$

Maximum likelihood estimate for θ is then given by

$$\begin{aligned}
&\underset{\theta}{\text{maximize}} \sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)};\theta) \\
&\equiv \underset{\theta}{\text{maximize}} \sum_{i=1}^{m} \log \frac{1}{1+\exp(-y^{(i)} \cdot \theta^T x^{(i)})} \\
&\equiv \underset{\theta}{\text{minimize}} \sum_{i=1}^{m} \log(1+\exp\left(-y^{(i)} \cdot h_\theta(x^{(i)})\right)
\end{aligned}$$

Softmax regression model

If instead Y takes on values in $\{1, \dots, k\}$, with

$$p(y = j|x; \theta) \propto \exp(\theta_j^T x)$$

Then

$$p(y = j|x; \theta) = \frac{\exp(\theta_j^T x)}{\sum_{l=1}^{k} \exp(\theta_l^T x)}$$

$$\log p(y = j|x; \theta) = \theta_j^T x - \log \sum_{l=1}^{k} \exp(\theta_l^T x)$$

$$\implies \underset{\theta}{\text{minimize}} \sum_{i=1}^{m} \left(\log \sum_{l=1}^{k} \exp(\theta_l^T x^{(i)}) - \theta_{y^{(i)}}^T x^{(i)} \right)$$

Least squares

In linear regression, assume

$$y = \theta^T x + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$$
$$\iff p(y|x;\theta) = \mathcal{N}(\theta^T x, \sigma^2)$$

Then the maximum likelihood estimate is given by

$$\underset{\theta}{\text{maximize}} \sum_{i=1}^{m} \log p\big(y^{(i)}\big|x^{(i)};\theta\big) \equiv \underset{\theta}{\text{minimize}} \sum_{i=1}^{m} \big(y^{(i)} - \theta^T x^{(i)}\big)^2$$

i.e., the least-squares loss function can be viewed as MLE under Gaussian errors

Other approaches possible too: absolute loss function can be viewed as MLE under Laplace errors

Outline

Probability in AI

Background on probability

Common distributions

Maximum likelihood estimation

Probabilistic graphical models

Probabilistic graphical models

Probabilistic graphical models are all about representing distributions

$$p(X)$$

where X represents some large *set* of random variables

Example: suppose $X \in \{0,1\}^n$ (n-dimensional random variable), would take $2^n - 1$ parameters to describe the full joint distribution

Graphical models offer a way to represent these same distributions more compactly, by exploiting *conditional independencies* in the distribution

Note: I'm going to use "probabilistic graphical model" and "Bayesian network" interchangeably, even though there are differences

Bayesian networks

A Bayesian network is defined by

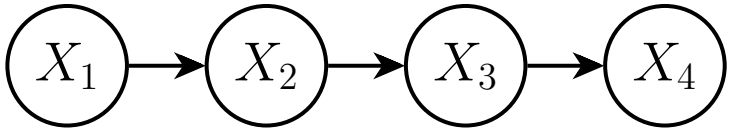
1. A directed acyclic graph, $G = \{V = \{X_1, \ldots, X_n\}, E\}$
2. A set of conditional distributions $p(X_i|\text{Parents}(X_i))$

Defines the joint probability distribution

$$p(X) = \prod_{i=1}^{n} p(X_i|\text{Parents}(X_i))$$

Equivalently: each node is conditionally independent of all non-descendants given its parents

Example Bayesian network



Conditional independencies let us simply the joint distribution:

$$p(X_1, X_2, X_3, X_4) = p(X_1)p(X_2|X_1)p(X_3|X_1, X_2)p(X_4|X_1, X_2, X_3)$$

$2^4 - 1 = 15$ parameters (assuming binary variables)

$$= p(X_1)p(X_2|X_1)p(X_3|X_2)p(X_4|X_3)$$

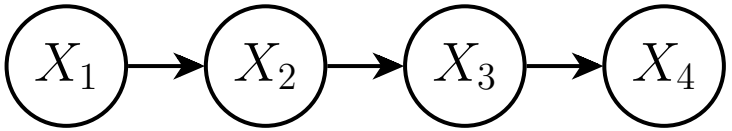
1 parameter

2 parameters

7 parameters

Generative model

Can also describe the probabilistic distribution as a sequential "story", this is called a *generative model*



$$
\begin{aligned}
X_1 &\sim \text{Bernoulli}\left(\phi^{(1)}\right) \\
X_2 | \; X_1 = x_1 &\sim \text{Bernoulli}\left(\phi^{(2)}_{x_1}\right) \\
X_3 | \; X_2 = x_2 &\sim \text{Bernoulli}\left(\phi^{(3)}_{x_2}\right) \\
X_4 | \; X_3 = x_3 &\sim \text{Bernoulli}\left(\phi^{(3)}_{x_3}\right)
\end{aligned}
$$

"First sample X_1 from a Bernoulli distribution with parameter $\phi^{(1)}$, then sample X_2 from a Bernoulli distribution with parameter $\phi^{(2)}_{x_1}$, where x_1 is the value we sampled for X_1, then sample X_3 from a Bernoulli …"

More general generative models

This notion of a “sequential story” (generative model) is extremely powerful for describing very general distributions

Naive Bayes:

$$Y \sim \text{Bernoulli}(\phi)$$
$$X_i | Y = y \sim \text{Categorical}\left(\phi_y^{(i)}\right)$$

Gaussian mixture model:

$$Z \sim \text{Categorical}(\phi)$$
$$X | Z = z \sim \mathcal{N}(\mu_z, \Sigma_z)$$

More general generative models

Linear regression:

$$Y|X = x \sim \mathcal{N}(\theta^T x, \sigma^2)$$

Changepoint model:

$$X \sim \text{Uniform}(0,1)$$

$$Y|X = x \sim \begin{cases} \mathcal{N}(\mu_1, \sigma^2) \text{ if } x < t \\ \mathcal{N}(\mu_2, \sigma^2) \text{ if } x \geq t \end{cases}$$

Latent Dirichlet Allocation: M documents, K topics, N_i words/document

$$\theta_i \sim \text{Dirichlet}(\alpha) \text{ (topic distributions per document)}$$
$$\phi_k \sim \text{Dirichlet}(\beta) \text{ (word distributions per topic)}$$
$$z_{i,j} \sim \text{Categorical}(\theta_i) \text{ (topic of } i\text{th word in document)}$$
$$w_{i,j} \sim \text{Categorical}(\phi_{z_i,j}) \text{ (}i\text{th word in document)}$$