Need for Sampling in ML

Sargur Srihari srihari@cedar.buffalo.edu

Topics

- 1. What is a sample from a known distribution?
- 2. A simple algorithm
- 3. Why is sampling needed in ML?
 - Probabilistic Graphical Models
 Intractability of Inference
 - 2. Bayesian Predictive Probability

What is a sample from a given distribution?

- Given a set of variables $x=\{x_1,...,x_d\}$
- A sample is an instantiation of an assignment to all variables

$$\mathbf{x}^{t} = \{x_1^t, \dots, x_d^t\}$$

 Each variable in a sample takes one possible allowable value in its domain according to a probability distribution defined over x

Univariate example

- Single variable x
- Domain of *x* is $\{x^0, ... x^{K-1}\}$
- Probability distribution is discrete P(x)
 - Example with K=4

	$P(x=x^{j})$
x^0	0.3
x^{l}	0.25
x^2	0.27
x^3	0.18

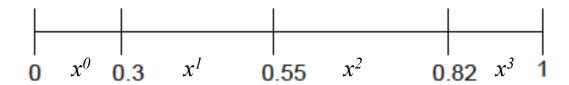
Examples of samples

$$x_1 = x^2$$
, $x_2 = x^0$, $x_3 = x^1$, $x_4 = x^0$,...

 x^0 repeats more often than others

Algorithm to generate univariate samples

- Domain of *x* is $\{x^0,...x^{K-1}\}$
- Probability distribution is discrete P(x)
- 1. Divide a real line [0,1] into K intervals such that width of interval is proportional to $P(x=x^j)$



- 2. Draw a random number $r \in [0,1]$
- 3. Determine region *j* in which *r* lies
- 4. Output x^j
 - -Random number r = 0.2929, x=?
 - -Random number r = 0.5209, x = ?

 $P(x=x^{j})$

0.3

0.25

0.27

0.18

 x^{0}

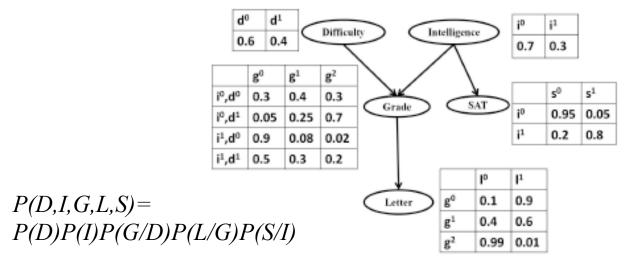
Why is sampling needed in ML?

- Inference is the task of answering probabilistic queries from model
- When exact inference is intractable, we need some form of approximation
 - True of probabilistic models of practical significance
- Samples can always be used to construct distributions
 - Inference methods based on numerical sampling are known as Monte Carlo techniques
 - Most situations will require evaluating expectations of unobserved variables, e.g., to make predictions
 - Rather than the posterior distribution

Probabilistic Graphical Models

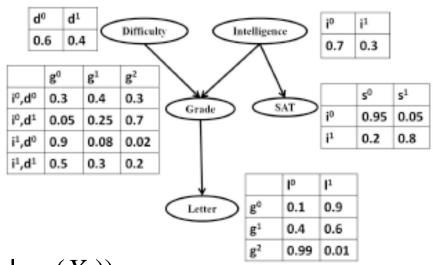
- PGMs (BNs and MNs) represent joint probability distributions over multiple variables
- Their use is to answer queries of interest
 - This is called "inference"
- What types of queries are there?
- Illustrate it with the "Student" Example

"Student" Bayesian Network



- Represents joint probability distribution over multiple variables
 - BNs represent them in terms of graphs and conditional probability distributions(CPDs)
 - Resulting in great savings in no of parameters needed

Joint distribution from Student BN



- CPDs: $P(X_i \mid pa(X_i))$
- Joint Distribution:

$$P(X) = P(X_1, ...X_n)$$

$$P(X) = \prod_{i=1}^{N} P(X_i \mid pa(X_i))$$

$$P(D, I, G, S, L) = P(D)P(I)P(G \mid D, I)P(S \mid I)P(L \mid G)$$

Query Types

1. Probability Queries

Given L and S give distribution of I

2. MAP Queries

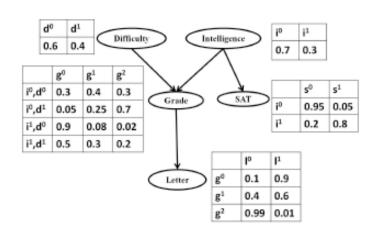
- Maximum a posteriori probability
- Also called MPE (Most Probable Explanation)
 - What is the most likely setting of D,I, G, S, L
- Marginal MAP Queries
 - When some variables are known

Probability Queries

- Most common type of query is a probability query
- Query has two parts
 - Evidence: a subset E of variables and their instantiation e
 - Query Variables: a subset Y of random variables in network
- Inference Task: P(Y|E=e)
 - Posterior probability distribution over values y of Y
 - Conditioned on the fact E=e
 - Can be viewed as Marginal over Y in distribution we obtain by conditioning on e
- Marginal Probability Estimation

$$P(Y = y_i | E = e) = \frac{P(Y = y_i, E = e)}{P(E = e)}$$

Example of Probability Query



$$P(Y = y_i \mid E = e) = \frac{P(Y = y_i, E = e)}{P(E = e)}$$
Posterior Marginal Probability of Evidence

- Posterior Marginal Estimation: $P(I=i^{l}|L=l^{0},S=s^{l})=?$
- Probability of Evidence: $P(L=l^0, s=s^1)=?$
 - Here we are asking for a specific probability rather than a full distribution

MAP Queries (Most Probable Explanation)

- Finding a high probability assignment to some subset of variables
- Most likely assignment to all non-evidence variables $W=\chi-Y$

```
MAP(W \mid e) = \arg \max_{w} P(w,e) Value of w for which P(w,e) is maximum
```

- Difference from probability query
 - Instead of a probability we get the most likely value for all remaining variables

Computing the Probability of Evidence

Probability Distribution of Evidence

$$P(L,S) = \sum_{D,I,G} P(D,I,G,L,S)$$
 Sum Rule of Probability
$$= \sum_{D,I,G} P(D)P(I)P(G \mid D,I)P(L \mid G)P(S \mid I)$$
 From the Graphical Model

Probability of Evidence

$$P(L = l^{0}, s = s^{1}) = \sum_{D,I,G} P(D)P(I)P(G \mid D,I)P(L = l^{0} \mid G)P(S = s^{1} \mid I)$$

More Generally

$$P(E=e) = \sum_{X \setminus E} \prod_{i=1}^{n} P(X_i \mid pa(X_i)) \mid_{E=e}$$

- An intractable problem
 - #P complete
 - Involves counting the number of possible solutions
- Tractable when tree-width is less than 25
 - Most real-world applications have higher tree-width

Approximate Solution using Sampling

Probability of Evidence

$$P(E=e) = \sum_{X \setminus E} \prod_{i=1}^{n} P(X_i \mid pa(X_i)) \mid_{E=e}$$

- Approximations are usually sufficient
 - When P(Y=y|E=e)=0.29292, approximation yields 0.3
- Sample the joint distribution over the variables
 - To determine how often E=e
 - E.g., use ancestral sampling to generate samples
 - This is the expected value of P(E=e)

Generative Models

- Many situations where we need to draw samples from a probability distribution
- Many methods of sampling exist
- Ancestral Sampling is relevant to graphical models
 - Given a graphical model we can specify how samples can be drawn from the joint or marginal distributions
 - Conditional can be drawn by setting the parent value to known value

Ancestral Sampling

- Start with lowest numbered node
- Draw a sample from the distribution $p(x_l)$ which we call \hat{x}_1
- Work through each of the nodes in order
 - For node n we draw a sample from conditional distribution $p(x_n|pa_n)$
 - Where parent variables are set to their sampled values
- Once final variable x_K is sampled
 - Achieved objective of obtaining a single sample from joint distribution
- To sample from marginal distribution
 - Sample from full distribution and discard unnecessary values
 - E.g., to draw from distribution $p(x_2,x_4)$ simply sample from full distribution, retain values $x_2^{\wedge},x_4^{\wedge}$ and discard remaining values $\{\hat{x}_{i\neq 2,4}\}$

Specifying Sampling Problem

• Joint distribution $p(x_1,...,x_K)$ over K variables which factorizes according to

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k \mid pa_k)$$

- Assume each node has a higher numbered node than its parents
- Goal is to draw a sample $\hat{x}_1, \hat{x}_2, ..., \hat{x}_K$
 - from this distribution

Numbering in Practical Applications

- Higher numbered variables
 - Correspond to terminal nodes of graph
 - Represent observations
- Lower numbered nodes
 - Correspond to latent variables
 - Latent variables allow
 - complicated distribution over observed variables to be represented in terms of simpler (typically exponential family) distributions

Sampling in Bayesian Prediction

- Given training data x and t and new test point x, goal is to predict value of t
 - i.e, wish to evaluate predictive distribution $p(t|x, \mathbf{x}, \mathbf{t})$
- Predictive distribution (where parameter has prior/ posteriors)

```
p(t \mid x, \mathbf{x}, \mathbf{t}) = \int p(t, \mathbf{w} \mid x, \mathbf{x}, \mathbf{t}) d\mathbf{w} Sum Rule (marginalizing over parameter w)

= \int p(t \mid x, \mathbf{w}, \mathbf{x}, \mathbf{t}) p(\mathbf{w} \mid x, \mathbf{x}, \mathbf{t}) d\mathbf{w} Prod. Rule (t given w & posterior of w)

= \int p(t \mid x, \mathbf{w}) p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}) d\mathbf{w} by eliminating unnecessary variables
```

With Gaussian noise: $p(t \mid x, \mathbf{w}) = N(t \mid y(x, \mathbf{w}), \beta^{-1})$ Posterior over parameters is obtained from prior by Bayes rule (likelihood x prior which is normalized)

 $\frac{p(\mathbf{t} \mid \mathbf{x}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{t} \mid \mathbf{x})}$ With Gaussian prior this is also Gaussian.

Convolution of two Gaussians is Gaussian giving a closed form solution. When distributions are complex the integration can be replaced by sampling

Bayesian Curve Fitting

 Posterior can be shown to be Gaussian

$$p(t | x, x, t) = N(t | m(x), s^{2}(x))$$

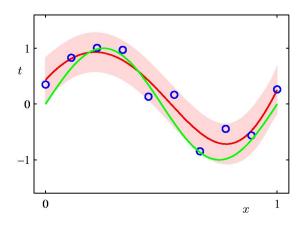
 Mean and Variance are dependent on x

$$m(x) = \beta \phi(x)^{T} S \sum_{n=1}^{N} \phi(x_{n}) t_{n}$$

$$s^{2}(x) = \beta^{-1} + \phi(x)^{T} S \phi(x)$$

$$S^{-1} = \alpha I + \beta \sum_{n=1}^{N} \phi(x_{n}) \phi(x)^{T}$$

$$\phi(x) \text{ has elements } \phi_{i}(x) = x^{i} \text{ for } i = 0,..M$$

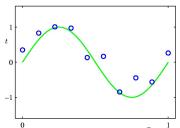


Predictive Distribution M=9 polynomial $a = 5 \times 10^{-3}$ b = 11.1Red curve is mean Red region is ± 1 std dev

Example: Bayesian Polynomial Regression

- Illustration of directed graph to describe probability distributions
- The polynomial regression problem

$$y(x,w) = \sum_{j=0}^{M} x^{j} w_{j}$$



- Probabilistic formulation with Random variables:
 - Vector of polynomial coefficients w

prior

- Observed data $t = (t_1,...,t_N)^T$ where t_n is a noise corrupted version of x_n
- Focusing on random variables, joint distribution is given by

$$p(t, \mathbf{w}) = p(\mathbf{w})p(\mathbf{t} \mid \mathbf{w})$$

$$= p(\mathbf{w})\prod_{n=1}^{N} p(t_n \mid \mathbf{w})$$
Since samples are independent

Conditional distributions

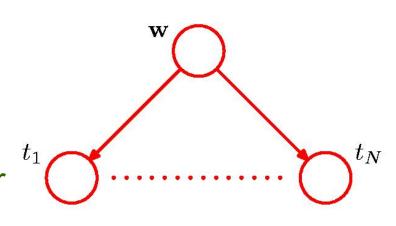
22

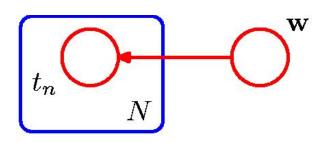
Graph for Polynomial Regression

Joint distribution is given by

$$p(t, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n \mid \mathbf{w})$$

- It is to be represented by a graphical model
 - Observing rhs introduce a node for each variable
 - Introduce associations
 - A link for each conditional distribution
- Equivalent plate representation





Srihari Machine Learning

Parameters in Bayesian Poly Regression

- Random variables:
 - Vector of polynomial coefficients w
 - Observed data $t = (t_1,...,t_N)^T$ where t_n is a noise corrupted version of x_n
- Additionally model contains:
 - Input data $\mathbf{x} = (x_1,...,x_N)^T$ Noise variance σ^2

 - Hyper-parameter lpha , which is the precision (inverse variance) of Gaussian prior

Parameters of model rather than random variables

Focusing on random variables, joint distribution is given by

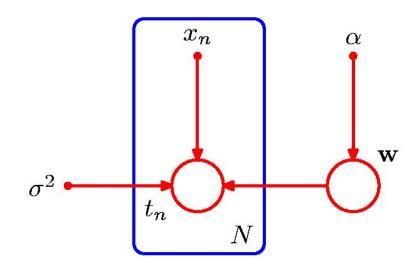
$$p(t, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n \mid \mathbf{w})$$
Gaussian prior Conditional distributions with precision α with variance σ^2

With Deterministic Parameters

- Sometimes useful to explicitly show parameters of model
- Same model with deterministic parameters

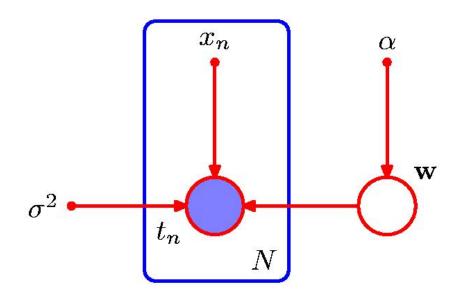
$$p(\mathbf{t}, \mathbf{w} | \mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2) = p(\mathbf{w} | \boldsymbol{\alpha}) \prod_{n=1}^{N} p(t_n | \mathbf{w}, x_n, \boldsymbol{\sigma}^2)$$

- Random variables are denoted by open circles
- Deterministic parameters by smaller (tiny) solid circles



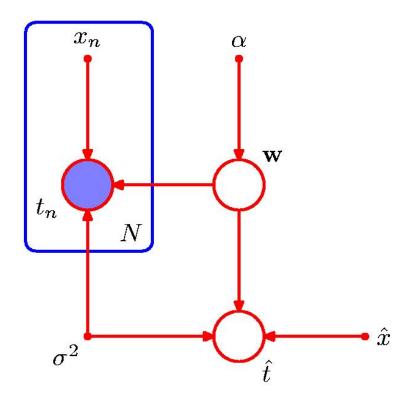
Observed and Latent Variables

- In ML some random variables are set to specific observed values
 - Observed variable nodes are shaded
- Variable w is not observed
 - It is latent or hidden
- Nodes t_n are shaded
 - to indicate corresponding variables are set to observed values



Including predictions

- Polynomial Regression model including input and predicted values
- Need to write out longer probability expression including
 - new input value $\hat{\chi}$ and model prediction $t^{\hat{}}$
- Joint distribution of all variables in this model



$$p(\hat{t}, \mathbf{t}, \mathbf{w} \mid \hat{\mathbf{x}}, \mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2) = p(\mathbf{w} \mid \boldsymbol{\alpha}) \prod_{n=1}^{N} p(t_n \mid \mathbf{w}, x_n, \boldsymbol{\sigma}^2) p(\hat{t} \mid \hat{x}, \mathbf{w}, \boldsymbol{\sigma}^2)$$

Predictive distribution

$$p(\hat{t} \mid \hat{x}, \mathbf{x}, \alpha, \sigma^2) \alpha \int p(\hat{t}, \mathbf{t}, w \mid \hat{x}, \mathbf{x}, \alpha, \sigma^2) d\mathbf{w}$$

