STA414
Lecture Notes

Yuchen Wang

January 5, 2020

Contents

1 Introduction

2 Introduction to Probabilistic Models

2.1 Overview of probabilistic models

In general, we have random variables $X = (X_1, \ldots, X_N)$ that are either *observed* or *unobserved*. Need a model that captures the relationship between these variables. The approach of probabilistic generative models is to relate all variables by a learned joint probability distribution $p_\theta(X_1, \ldots, X_N)$. We assume there is a true joint p_*, which we are trying to learn with a model p_θ.
Assume we have the joint probability $p(X, C, Y)$

Regression

$$p(Y|X) = \frac{p(X,Y)}{p(X)} = \frac{p(X,Y)}{\int p(X,Y)\, dY}$$

Classification / Clustering

$$p(C|X) = \frac{p(X,C)}{\sum_C p(X,C)}$$

Assigning the class label:

1. $c^* = \arg\max_c p(C = c|X)$
2. Sample the class assignment from our distribution, $c^* \sim p(C|X)$
3. Output the class assignment along with its density under our distribution $(c^*, p(C = c^*|X))$. Can inform us of the model's uncertainty or confidence of the prediction.

Latent/hidden Variables Variables which are never observed in the dataset.

Operations on Probabilistic Models

- **Generate Data**
- **Estimate Likelihood**
- **Inference**: Compute expected value of some variables given others which are either observed or marginalized.
- **Learning**: Set the parameters of the joint distribution given some observed data to maximize the probability of the observed data.

Goals of joint distributions

1. Facilitate **efficient computation** of marginal and conditional distributions
2. Have compact representation so the size of the parameterization scales well for joint distributions over many variables.

Joint Dimensionality Suppose n is our number of variables and k our states. The dimensionality of our parameters then becomes k^n

2.2 Sufficient statistics

Definition 2.1 (Statistic and Sufficient statistic)**.** A statistic is a (possibly vector valued) deterministic function of a (set of) random variable(s). A sufficient statistic is a statistic that conveys exactly the same information about the data generating process that created the data as the entire data itself. Formally, we say that $T(X)$ is a sufficient statistic for X if

$$T(x^{(1)}) = T(x^{(2)}) \implies L(\theta; x^{(1)}) = L(\theta; x^{(2)}) \quad \forall \theta$$

where L is the likelihood function.
Alternatively,

$$P(\theta|T(X)) = P(\theta|X)$$

Equivalently (by the Neyman factorization theorem) we can write

$$P(\theta|T(X)) = h(x, T(x))g(T(x), \theta)$$

An example is the exponential family

$$p(x|\eta) = h(x)\exp\left\{\eta^T T(x) - g(\eta)\right\}$$

or, equivalently

$$p(x|\eta) = h(x)g(\eta)\exp\left\{\eta^T T(x)\right\}$$

Example 2.1 (Bernoulli Trials)**.** We observe N iid coin flips.
Model: $p(H) = \theta, P(T) = 1 - \theta$
Likelihood: $l(\theta; D) = \log\theta \sum_n x^{(n)} + \log(1-\theta)\sum_n (1 - x^{(n)})$
Notice that our likelihood depends on $\sum_n x^{(n)}$.
$\implies$ If we know this summary statistic $T(x) = \sum_n x^{(n)}$, then we know everything that is useful from our sample todo inference.

$$l(\theta; D) = T(X)\log\theta + (N - T(X))\log(1-\theta)$$

Then we take the derivative and set it to 0 to find the maximum

$$\Rightarrow \frac{\partial \ell}{\partial \theta} = \frac{T(X)}{\theta} - \frac{N - T(X)}{1-\theta}$$
$$\Rightarrow \hat{\theta} = \frac{T(X)}{N}$$

This is our maximum likelihood estimation of the parameters $\theta, \theta^{\star}_{MLE}$.

Example 2.2 (Multinomial)**.** We observe M iid die rolls (K-sided).
Model: $p(k) = \theta_k, \sum_k \theta_k = 1$
Likelihood: $l(\theta; D) = \sum_k N_k \log\theta_k$
Take derivatives and set to zero (enforcing $\sum \theta_k = 1$):

$$\frac{\partial \ell}{\partial \theta_k} = \frac{N_k}{\theta_k} - M$$
$$\Rightarrow \theta^*_k = \frac{N_k}{M}$$

sufficient statistics: number of each type

Remark 2.1 (Sufficient statistics are sums)**.** For all exponential family models, sufficient statistics are the average natural parameters.

3 Directed Graphical Models

Notation 3.1. The joint distribution of N random variables can be computed by the chain rule

$$p\left(x_{1,\ldots,N}\right) = p\left(x_1\right)p\left(x_2|x_1\right)p\left(x_3|x_2,x_1\right)\ldots p\left(x_n|x_{n-1:1}\right)$$

this is true for any joint distribution over any random variables (assuming full dependence between variables). More formally, in probability the chain rule for two random variables is

$$p(x,y) = p(x|y)p(y)$$

and for N random variables

$$p\left(\cap_{i=1}^{N} x_i\right) = \prod_{j=1}^{N} p\left(x_j | \bigcap_{k=1}^{j-1} x_k\right)$$

We can represent a model $p(x_i, x_{\pi_i}) = p(x_{\pi_i})p(x_i|x_{\pi_i})$ as a graph where nodes represent random variables and arrows mean "conditioned on".
We can simplify the model by building in our assumptions about the conditional probabilities.

3.1 Directed acyclic graphical models (DAGM)

Definition 3.1. A directed acyclic graphical model over N random variables looks like

$$p(x_{1,\ldots,N}) = \Pi_i^N p(x_i|x_{\pi_i})$$

where $x)i$ is a random variable or a node in the graphical model and x_{π_i} are the parents of this node.
In other words, the joint distribution factors into a product of conditional distributions.
Missing edges imply conditional independente.

Remark 3.1. We are conditioning on parent nodes as opposed to every node. Therefore, the model that represents this distribution is exponential in the fan-in of each node (the number of nodes in the parent set), instead of in N.

Definition 3.2 (D-Separation)**.** D-separation, or directed-separation is a notion of connectedness in DAGMs in which two (sets of) variables may or may not connected conditioned on a third (set of) variable(s).
D-connection implies conditional dependence and d-separation implies conditional independence.