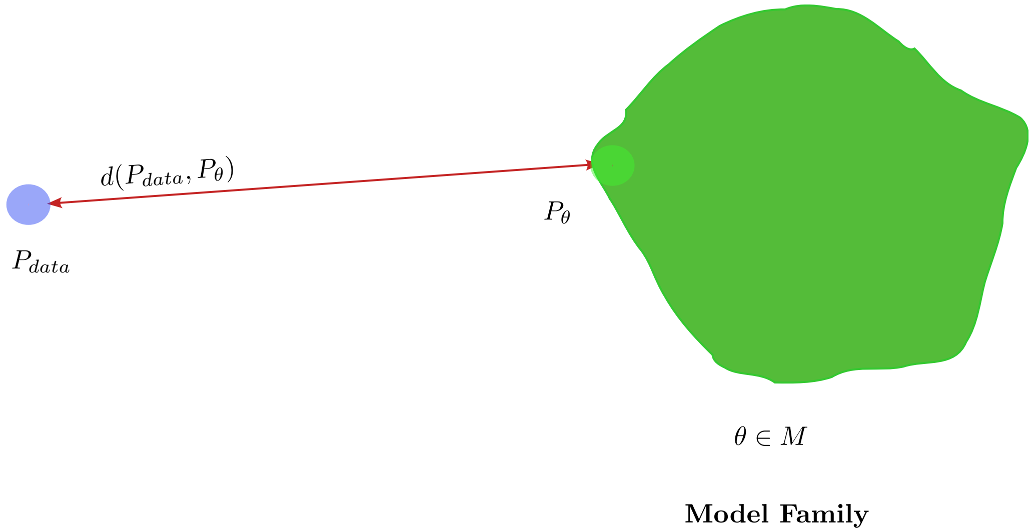
CS236 Lecture 2

May 8, 2024

1 The problem space



- Over on the left, the blue represents the real world system from which we have data samples. It could be a set of images of dogs, for example. So $x_i \sim P_{data}$
- The green area is the space of possibilities of probability distributions that are parameterised by θ.
- We need to define the notion of distance or loss function d.

We want to learn a probability distribution $p(x)$ over images x such that:

- **Generation:** If we sample $x_{new} \sim p(x)$, x_{new} should look like a dog (*sampling*)
- **Density estimation:** $p(x)$ should be high if x looks like a dog, and low otherwise (*anomaly detection*)
- **Unsupervised representation learning:** We should be able to learn what these images have in common, e.g. ears, tails, etc. (*features*)

2 How do you represent the probability distribution $p(x)$?

For low dimensional data it would be straightforward. In the simplest case, binary distribution biased coin flip (Heads or Tails) is modelled by the Bernoulli Distribution.

- $D = \{Heads, Tails\}$
- Specify $P(X = Heads) = p$. Then $P(X = Tails) = 1 - p$
- Write $X \sim Ber(p)$
- Sampling: flip a (biased) coin

Extending this, we can have a Categorical distribution: a biased m-sided dice.

- $D = \{1, \cdots, m\}$
- $P(Y = i) = p_i$ such that $\sum p_i = 1$
- Write $Y \sim Cat(p_1, \cdots, p_m)$
- Sampling: roll a biased die

2.1 Images

When we consider images, they are comprised of pixels. And each pixel can be considered three random variables, R, G, B each of which is taken from $\{0, \cdots, 255\}$ Sampling from the joint distribution $(r, g, b) \sim p(R, G, B)$ randomly generates a colour for the pixel. The number of parameters to represent this space is $256 * 256 * 256 - 1$. We have the -1 because the probabilities we know have to add up to one, so the last parameter is implicitly known by subracting from 1 the sum of all the other parameters. The problem is that for modest images, the number of parameters is too large. E.g. 23 x 23 pixels black and white for representing a handwritten digit means $2^{529} - 1$ parameters. We need to have some simplifying assumptions.

3 Structure through independence

If $X_1, \cdots, X_n$ are independent, then

$$p(x_1, \cdots, x_n) = p(x_1)p(x_2)\cdots p(x_n) \tag{1}$$

With this assumption how many states are needed for the images? It is 2^n where n is the number of pixels. If the images are black and white, $|Val(X_i)|= 2$. So 2^n entries can be described by just n numbers.

In reality the independence assumption is too strong. Because for digits handwritten, you won't get random scattering of pixels. They will be lit up in a localised area representing the stroke of the pen.

What we do is instead to make conditional independence assumptions.

1. **Chain rule** Let $S_1, \cdots S_n$ be events, $p(S_i) > 0$.
   $p(S_1 \cap S_2 \cap \cdots \cap S_n) = p(S_1)p(S_2|S_1)\cdots p(S_n|S_1 \cap \cdots \cap S_{n-1})$

2. **Bayes' rule** Let S_1, S_2 be events, $p(S_1) > 0$ and $p(S_2) > 0$.
   $p(S_1|S_2) = \frac{p(S_1 \cap S_2)}{p(S_2)} = \frac{p(S_2|S_1)p(S_1)}{p(S_2)}$

In the Chain Rule, the joint distributions on the left are hard to know but the conditional probabilities (aka marginal probabilities) terms on the right are easier to obtain. This is why the Chain Rule helps us.

Using Chain Rule

$$p(x_1, \cdots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \cdots p(x_n|x_1, \cdots, x_{n-1}) \tag{2}$$

This is the kind of factorisation used in autoregressive models. In fact, any collection of random variables can be factored in this way.

How many parameters do we now need?

$1 + 2 + 4 + \cdots + 2^{n-1} = 2^n - 1$

The reason why this is a geometric series is because with x_i being 0 or 1 then x_1, x_2 means four possibilities so $p(x_3|x_1, x_2)$ requires 4 parameters. So we have not yet made progress on simplifying our computation.

Now let us suppose that for X_{i+1} where X_{i+1} represents the weather on day $i+1$ is dependent on only the weather on day i and independent of the weather on prior days $1, \cdots, i-1$. Then we can write this as $X_{i+1} \perp X_1, \cdots X_{i-1} \mid X_i$

This allows us to simplify because for example the third term $p(x_3|x_1, x_2)$ becomes $p(x_3|x_2)$. This is true for the second to the $(n-1)th$ term. Now we are conditioning on at most one variable. Each of those terms requires two parameters; one for the $x_i = 0$ case and one for the $x_i = 1$ case. The first term requires just one parameter since it is one variable alone.

How many parameters do we now need? $1 + \underbrace{2 + 2 + \cdots + 2}_{n-1} = 2n - 1$

This is a exponential improvement but the predictions may not be good as just using the prior event alone is not enough context for good prediction, such as predicting the next word in a sentence.

4 Bayes' Network

Here we use conditional parameterisation instead of joint parameterisation. For each random variable X_i specify $p(x_i|\boldsymbol{x}_{\boldsymbol{A_i}})$ for set $\boldsymbol{X}_{\boldsymbol{A_i}}$ of random variables.

Then the joint parameterisation is $p(x_1, \cdots, x_n) = \prod_i p(x_i|\boldsymbol{x}_{\boldsymbol{A_i}}))$

We define a **Bayesian network** as a *directed* **acyclic** graph (DAG)

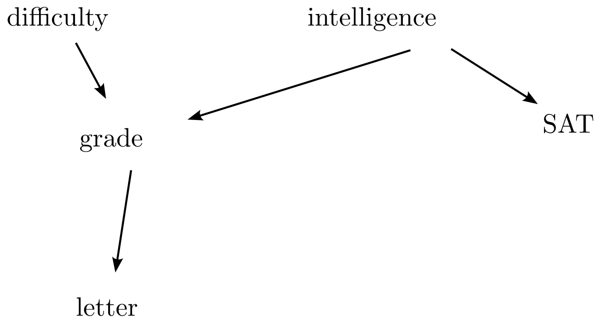
$$G = (V, E) \tag{3}$$

- One node $i \in V$ for each random variable X_i
- One node for each pixel in images, one node for each word in sentences
- One conditional probability distribution (CPD) per node, $p(x_i|\boldsymbol{x}_{Pa(i)})$, specifying the variable's probability conditioned on its parents' values

$$p(x_1, \cdots, x_n) = \prod_{i \in V} p(x_i|\boldsymbol{x}_{Pa(i)})) \tag{4}$$

- We claim that $p(x_1, \cdots, x_n)$ is a valid probability distribution because of ordering implied by DAG.
- **Economical representation:** exponential in $|Pa(i)|$ not $|V|$

5 Course Grade Example



Using the terminology d, i, g, s, l for difficulty, intelligence, grade, SAT, and letter respectively, the joint distribution is:

$$p(x_1, \cdots, x_n) = \prod_{i \in V} p(x_i | \boldsymbol{x}_{Pa(i)}) \tag{5}$$

$$p(d, i, g, s, l) = p(d)p(i)p(g|i, d)p(s|i)p(l|g) \tag{6}$$

However, it is always true, by the chain rule, *any* distribution can be written as

$$p(d, i, g, s, l) = p(d)p(i|d)p(s|i, d, g)p(l|g, d, i, s) \tag{7}$$

Therefore, we are assuming the following additional independencies:

- $D \perp I$
- $S \perp \{D, G\} | I$
- $L \perp \{I, D, S\} | G$

6 Discriminative model

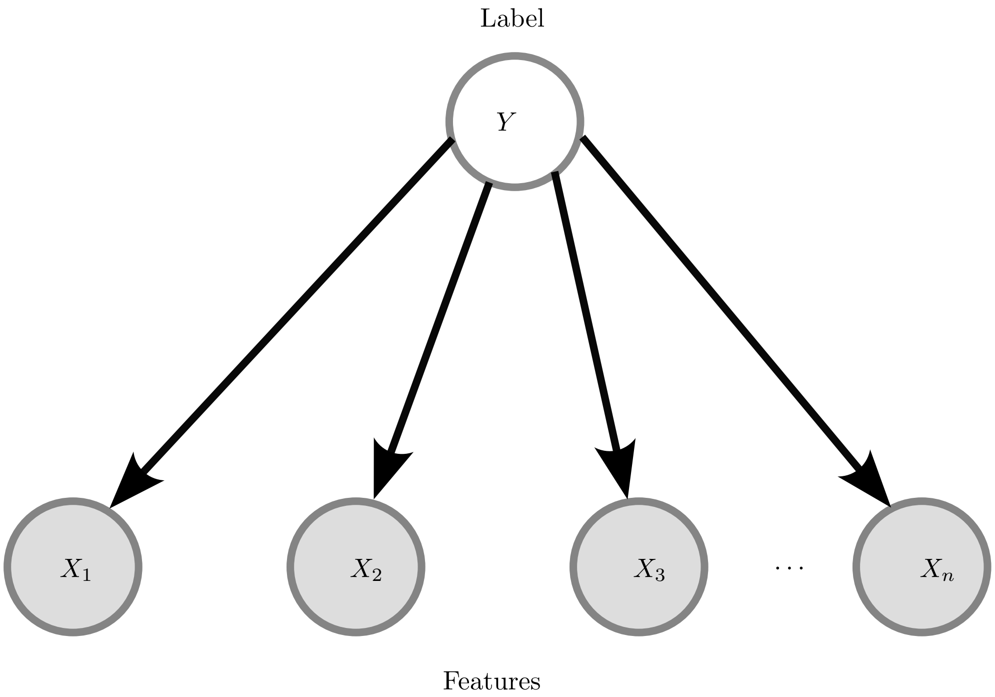
What is the difference between the discriminative model approach and the generative approach? We will look at the discriminative case here for e-mail spam detection.

Classify e-mails as spam ($Y = 1$) or not spam ($Y = 0$)

- Let $1 : n$ index the words in our vocabulary (e.g. English)

- $X_i = 1$ if word i appears in an e-mail, and 0 otherwise
- E-mails are drawn according to some distribution $p(Y, X_1, \cdots, X_n)$

The naive Bayes for this setup is to make the words conditionally independent given Y:



Then

$$p(y, x_1, \cdots, x_n) = p(y) \prod_{i=1}^{n} p(x_i \mid y) \tag{8}$$

We **estimate** the parameters from training data. We **predict** with Bayes rule:

$$p(Y = 1 \mid x_1, \cdots, x_n) = \frac{p(Y = 1) \prod_{i=1}^{n} p(x_i \mid Y = 1)}{\sum_{y=\{0,1\}} p(Y = y) \prod_{i=1}^{n} p(x_i \mid Y = y)} \tag{9}$$