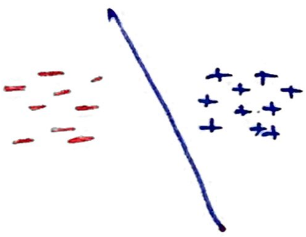
Unit 4 - Lecture 15: Generative Models

Objectives:

- Understand what Generative Models are and how they work
- Understand estimation and prediction phases of generative models
- Derive a relation connecting generative and discriminative models
- Derive Maximum Likelihood Estimates (MLE) for multinomial and Gaussian generative models.

Generative vs Discriminative models:

Whenever we talked about classification, the pictures that you had in mind is that your classifier had a training data, let's say positives and negatives instances. The job of the discriminative model was to find a separator that discriminates between these points.



The approach that we will be taking is actually trying to understand what is the structure of these ⊕ and ⊖ classes.

If I can understand in some probabilistic term, the structure of positive and negative examples maybe I can do discrimination in a different way that we've done when we're primarily looking at the separator.

Two questions to ask about any generative model:

1. how we estimate this model? Estimation question
2. how we actually do prediction? Prediction question

Given our training data, we will fit probability distribution for the negative class and for the positive class. And by comparing these probabilities, we can actually induce what is the right label.

Simple Multinomial Generative model:

We will talk about multinomial and in your head, you can think about texts documents. and what our models will do, they will generate documents. Some parameters:

$$p(w \mid \theta) = \theta_w$$ ← captures what's the likelihood of me selecting, generating certain words given all the possibilities.

w: a word; θ: parameters of a model

Selecting words independently from each other.

$\theta_w \geq 0$, $\sum_{w \in W} \theta_w = 1$

likelihood function:

let's say we have θ_w, how do I compute the likelihood of generating a document?

$$P(D|\theta) = \prod_{i=1}^{n} \underbrace{\theta_{w_i}}_{\text{probability of every word in our document}}$$

we have our document

$$P(D|\theta) = \underbrace{\prod_{w \in W}}_{\text{the words in our vocabulary}} \theta_w^{count(w)}$$

let's take an example: We have two words in our vocabulary

$$W = \{cat, dog\}$$

model 1 which takes θ, $\theta_{cat} = 0.3$, $\theta_{dog} = 0.7$

model 2 which takes θ', $\theta'_{cat} = 0.9$, $\theta'_{dog} = 0.1$

$D = \{cat, cat, dog\}$ we can compute the likelihood of these documents generated by 1 and 2 model

$$P(D|\theta) = (0.3)^2 \times (0.7)$$

$$P(D|\theta') = (0.9)^2 \times (0.1)$$

How can we utilize our training data to find the best parameters?

We will use Maximum likelihood, and will make assumption, that the best parameters are the parameters which give the highest likelihood to our data.

Find θ's which maximize $P(D|\theta) = \max_{\theta} \prod_{w \in W} \theta_w^{count(w)}$

$$\log \prod_{w \in W} \theta_w^{count(w)} = \sum_{w \in W} count(w) \log \theta_w$$

$W = \{\underset{cat}{0}, \underset{dog}{1}\}$

$\theta_{cat} = \theta_0 = \theta$

$\theta_{dog} = \theta_1 = 1 - \theta$

$$= \frac{\partial}{\partial \theta} count(0) . \log(\theta) + count(1) . \log(1 - \theta)$$

$$\Longrightarrow \frac{count(0)}{\theta} - \frac{count(1)}{1 - \theta} = 0$$

$$(1 - \theta)\, count(0) - \theta\, count(1) = 0$$

$$\hat{\theta} = \frac{count(0)}{count(1) + count(0)}$$

Prediction:

MLE for Multinomial Distribution: $\hat{\theta}_w = \dfrac{\text{count}(w)}{\sum_{w \in W} \text{count}(w')}$

Using the estimation techniques described earlier, we can find θ^+ and θ^- that they will give the Highest likelihood to the points.

The question is, if I give you a new document, how do you know to which class it belongs?

I could look at the likelihood that this document was generated by plus side and minus side.

$$\log \frac{P(D|\theta^+)}{P(D|\theta^-)}$$

For now let's assume that the prior likelihood of the ⊖ and ⊕ class are exactly the same.

In this case: $$\log \frac{P(D|\theta^+)}{P(D|\theta^-)} = \begin{cases} \geq 0 \rightarrow + \\ < 0 \rightarrow - \end{cases}$$

→ class conditional distribution.

$$= \log P(D|\theta^+) - \log P(D|\theta^-)$$

$$= \log \prod_{w \in W} \theta_w^{+\,\text{count}(w)} - \log \prod_{w \in W} \theta_w^{-\,\text{count}(w)}$$

$$= \sum_{w \in W} \text{count}(w) \log \theta_w^+ - \sum_{w \in W} \text{count}(w) \log \theta_w^-$$

$$= \sum_{w \in W} \text{count}(w) \underbrace{\log \frac{\theta_w^+}{\theta_w^-}}_{\tilde{\theta}_w}$$

$$= \sum_{w \in W} \text{count}(w) \cdot \tilde{\theta}_w$$

Prior, Posterior and likelihood:

Sometimes we migh have some prior knowledge and we want to take advantage of it.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$ Bayesian Rule

$$\underbrace{P(y=+|D)}_{\text{posterior}} = \frac{P(D|\theta^+)\ \overbrace{P(y=+)}^{\text{Prior}}}{P(D)}$$

What is the likelihood that I'm going to assigned to document D it's label y which is ⊕?

$$\log \frac{P(y=+|D)}{P(y=-|D)} = \log \frac{P(D|\theta^+) \cdot P(y=+)}{P(D|\theta^-) \cdot P(y=-)}$$

$$\log \frac{P(D|\theta^+)\,P(y=+)}{P(D|\theta^-)\,P(y=-)} = \log \frac{P(D|\theta^+)}{P(D|\theta^-)} + \underbrace{\log \frac{P(y=+)}{P(y=-)}}_{\tilde{\theta}_0}$$

$$= \sum_{w \in W} \text{count}(w) . \tilde{\theta}_w + \tilde{\theta}_0$$ We translate it to a linear classifier

Gaussian Generative models:

Now we are going to vectors in $\mathbb{R}^d$ and find distribution that is a very natural fit for describing this data. $X \in \mathbb{R}^d$

We will describe these points by two parameters:

1. Where is the center of the cloud? μ
2. How dispersed is this cloud? σ^2

$$P(X|y\sigma^2) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2} \|x-\mu\|^2\right)$$

Multivariate Gaussian Random vector:

A Random vector $X = (X^{(1)}, \ldots X^{(d)})^T$ is a Gaussian vector, or multivariate Gaussian or normal variable, if any linear combination of its components is a (univariate) Gaussian variable or a constant (a "Gaussian" variable with zero variance) i.e, if $\alpha^T X$ is (univariate) Gaussian or constant for any constant non-zero vector $\alpha \in \mathbb{R}^d$

The distribution of X, the d-dimensional Gaussian or normal distribution, is completely specified by the vector mean $\mu = E[X] = (E[X^{(1)}], \ldots, E[X^{(d)}])^T$ and the $d \times d$ covariance matrix Σ. if Σ is invertible, then the pdf of X is:

$$f_X(X) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \qquad X \in \mathbb{R}^d$$

Where $\det(\Sigma)$ is the determinant of the Σ, which is positive when Σ is invertible

if $\mu = 0$ and Σ is the identity matrix, then X is called a Standard normal random vector

MLE for Gaussian Distribution:

In this problem we will derive the maximum likelihood estimates for a Gaussian model

$$f_X(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{\|x-\mu\|^2}{2\sigma^2}}$$

Let $S_m = \{X^{(1)}, X^{(2)}, \ldots X^{(m)}\}$ be i.i.d random variables following a Gaussian distribution with mean μ and variance σ^2, then their joint probability density function is given by:

$$\prod_{t=1}^{m} P(x^{(t)} \mid \mu, \sigma^2) = \prod_{t=1}^{m} \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{\|x-\mu\|^2}{2\sigma^2}}$$

Taking logarithm of the above function, we get:

$$\log P(S_m \mid \mu, \sigma^2) = \log\left(\prod_{t=1}^{m} \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{\|x-\mu\|^2}{2\sigma^2}}\right)$$

$$= \sum_{t=1}^{m} \log \frac{1}{(2\pi\sigma^2)^{d/2}} + \sum_{t=1}^{m} \log e^{-\frac{\|x^t-\mu\|^2}{2\sigma^2}}$$

$$= \sum_{t=1}^{m} -\frac{d}{2} \log(2\pi\sigma^2) + \sum_{t=1}^{m} \log e^{-\frac{\|x-\mu\|^2}{2\sigma^2}}$$

$$= -\frac{md}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=1}^{m} \|x^{(t)}-\mu\|^2$$

$$\frac{\partial \log P(S_m \mid \mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{t=1}^{m} (x^{(t)} - \mu) = 0$$

$$\hat{\mu} = \frac{\sum_{t=1}^{m} x^{(t)}}{m}$$

$$\frac{\partial \log P(S_m \mid \mu, \sigma^2)}{\partial \sigma^2} = -\frac{md}{2\sigma^2} + \frac{\sum_{t=1}^{m} \|x^{(t)}-\mu\|^2}{2(\sigma^2)^2}$$

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{m} \|x^{(t)}-\mu\|^2}{md}$$