Intro. Problem Statement

Denis Derkach, Maksim Artemev, Artem Ryzhikov

CS HSE faculty, spring 2020

Contents

Intuition

Generating examples

- You have some amount of measurements:

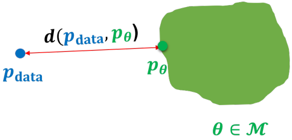
$$\{1; 0; 1; 1; 0; 1; 0; 1\}$$

- Can you write out an element of this set?
- How did you do this?
- How would you make your action more precise?

What do we want to do

- We have a sample of data objects.
- We want to have:
  - way to sample new objects $x \sim p_\theta(x)$ that look similar to given ones;
  - way to estimate $p_\theta(x)$;
  - way to learn common features in unsupervised manner.

$$\mathbf{x}^{(j)} \sim p_{\text{data}}$$
$$j = 1, 2, \ldots, |\mathcal{D}|$$



Model family

Here and later from S. Ermon and A. Grover

Probability

What is a Probability?

- The quality or state of being probable; the extent to which something is likely to happen or be the case. (Oxford dictionaries).
- Generally, can be understood without any knowledge of mathematics.
- However, mathematics is quite essential to understand the subject.

see Goodfellow et al. Deep Learning Book Part I Chap 3

Kolmogorov axioms

For event space $\mathcal{F}$ with given function $\mathbb{P}$:

- The probability of event $A \in \mathcal{F}$ is assigned a non-negative real number $\mathbb{P}(A)$, which is called the probability of A.
- The probability of at least one event from $\mathcal{F}$ to occur: $\mathbb{P}(\mathcal{F}) = 1$.
  - (*) The probability of an empty set of events is $\mathbb{P}(\emptyset) = 0$.
- If $X_1 \in \mathcal{F}$ and $X_2 \in \mathcal{F}$ are mutually exclusive, than $\mathbb{P}(X_1 + X_2) = \mathbb{P}(X_1) + \mathbb{P}(X_2)$ (also for any countable number of events).

Generally, other sets of axioms are possible. The main question stays: how we interpret what stays behind our probabilities.

Some Properties of Probability

- Joint event probabilities $P(A \text{ or } B)$ and $P(A \text{ and } B)$:

$$\mathbb{P}(A \text{ or } B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \text{ and } B)$$

  .

- Full probability:

$$\mathbb{P}(A) = \sum_n \mathbb{P}(A \text{ and } B_n)\mathbb{P}(B_n),$$

  where the whole space can be partitioned into a set of B_n,

- Conditional probability, $\mathbb{P}(A|B)$, means the probability that A is true, given that B is true.

Bayes Theorem

- For a joint probability:

$$\mathbb{P}(A \text{ and } B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$$

.

- Which implies:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

.

- Using Full probability:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|notB)\mathbb{P}(notB)}$$

Example for Bayes Theorem

Suppose we have a particle ID detector designed to identify particle of type K, with the property that if a K hits the detector, the probability that it will produce a positive pulse (T^+) is 0.9:

$$P(T^+|K) = 0.9[90\% \text{ acceptance}]$$

and 1% if a noise particle goes through:

$$P(T^+|notK) = 0.01[1\% \text{ background}]$$

Now a particle gives a positive pulse. What is the probability that it is a K?

Example for Bayes Theorem

The answer by Bayes Theorem:

$$\mathbb{P}(K|T^+) = \frac{\mathbb{P}(T^+|K)\mathbb{P}(K)}{\mathbb{P}(T^+|K)\mathbb{P}(K) + \mathbb{P}(T^+|notK)\mathbb{P}(notK)}$$

. In other words, all depends on the $\mathbb{P}(K)$.

K in beam	$\mathbb{P}(K) = 1\%$	$\mathbb{P}(K) = 10^{-6}\%$
$\mathbb{P}(K\|T+)$	0.48	10^{-4}
$\mathbb{P}(K\|T-)$	0.01	10^{-7}

- Bayes theorem can be used to easily solve the problem.
- This detector is not very useful if $\mathbb{P}(K)$ is small.
- No interpretation of $\mathbb{P}$ is given (you can be Bayesian or Frequentist).

Random Variable

A Random Variable is a variable which will take different values if the experiment is repeated.
These values are unpredictable except that we know in probability:

$$\mathbb{P}(data|parameters),$$

provided any unknowns in the parameters are given some assumed values.

Probability density function

When the data are continuous, the probability of a random variable ξ, $\mathbb{P}$, can be rewritten as Probability Density Function, or PDF:

$$p_{\xi|parameters}(x)dx = \mathbb{P}(\xi \in [x; x + dx]|parameters).$$

We normally write something like:

$$\mathbb{P}(\xi|parameters) = p(x; parameters).$$

NB: the same can be written for discrete random variables and is called probability mass function.

Basic discrete distributions

- Bernoulli distribution: (biased) coin flip
  - $D = \{Heads, Tails\}$
  - Specify $P(X = Heads) = p$. Then $P(X = Tails) = 1 - p$.
  - Write: $X \sim Ber(p)$
  - Sampling: flip a (biased) coin
- Categorical distribution: (biased) m-sided dice
  - $D = \{1, \cdots, m\}$
  - Specify $P(Y = i) = p_i$, such that $\sum p_i = 1$
  - Write: $Y \sim Cat(p_1, \cdots, p_m)$
  - Sampling: roll a (biased) die

Cumulative Density Function (CDF)

Definition

The cumulative distribution function (cdf) is the probability that the variable takes a value less than or equal to x. That is:

$$F(x) = \mathbb{P}[X \leq x].$$

Basic Characteristics of PDF

If we have a PDF $p_\xi(x)$ of a random variable ξ.

- Expectation:

$$\mathbb{E}(\xi) = \int x p_\xi dx,$$

- Variance:

$$\mathbb{V}ar_\xi(\xi) = \mathbb{E}_\xi\left[(\xi - \mathbb{E}_\xi(\xi))^2\right]$$

  ,

- Higher central momenta:

$$\mu_\xi^k = \mathbb{E}_\xi\left[(\xi - \mathbb{E}_\xi \xi)^3\right],$$

Properties of Expectation and Variance

- Expectation
  - $\mathbb{E}(c) = c$;
  - $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$;
  - For independent X and Y: $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.
- Variance
  - $\mathbb{V}ar(c) = 0$;
  - $\mathbb{V}ar(X) \geq 0$;
  - $\mathbb{V}ar(X + c) = \mathbb{V}ar(X)$;
  - $\mathbb{V}ar(cX) = c^2\mathbb{V}ar(X)$.

Multidimensional distributions

We often encounter situations where we have to analyze several random variables at once. In this case, we need to analyze a more complex entity, the multidimensional PDF $\mathbb{P}(\xi_1 \leq x_1, \ldots, \xi_n \leq x_n)$ for a random vector $\xi = (\xi_1, \ldots, \xi_n)$.

Independence of random variables

Definition

Let random variables X and Y have a joint density $p(x, y)$. X and Y will be called independent if

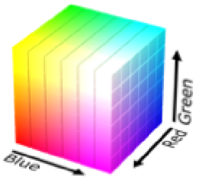
$$p(x, y) = p(x) \cdot p(y).$$

NB: Generally, it is more appropriate to use mutual information concept.

Example of joint distribution

Modeling a single pixel's color. Three discrete random variables:

- Red Channel R. $\mathrm{Val}(R) = \{0, \cdots, 255\}$
- Green Channel G. $\mathrm{Val}(G) = \{0, \cdots, 255\}$
- Blue Channel B. $\mathrm{Val}(B) = \{0, \cdots, 255\}$



Sampling from the joint distribution $(r, g, b) \sim p(R, G, B)$ randomly generates a color for the pixel. How many parameters do we need to specify the joint distribution $p(R = r, G = g, B = b)$?

$$256 * 256 * 256 - 1$$

Example of joint distribution

- Suppose $X_1, \ldots, X_n$ are binary (Bernoulli) random variables, i.e., $\text{Val}(X_i) = \{0, 1\} = \{\text{Black}, \text{White}\}$.
- How many possible states?

$$\underbrace{2 \times 2 \times \cdots \times 2}_{n \text{ times}} = 2^n$$

- Sampling from $p(x_1, \ldots, x_n)$ generates an image
- How many parameters to specify the joint distribution $p(x_1, \ldots, x_n)$ over n binary pixels?

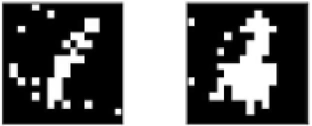
$$2^n - 1$$

Independent distribution

- If $X_1, \ldots, X_n$ are independent, then

$$p(x_1, \ldots, x_n) = p(x_1)p(x_2)\cdots p(x_n)$$

- How many possible states? 2^n
- How many parameters to specify the joint distribution $p(x_1, \ldots, x_n)$?
  - How many to specify the marginal distribution $p(x_1)$? 1
- 2^n **entries can be described by just** n **numbers** (if $|\mathrm{Val}(X_i)| = 2$)!
- Independence assumption is too strong. Model not likely to be useful
  - For example, each pixel chosen independently when we sample from it.



Conditional Independence

Definition

Two events A, B are conditionally independent given event C if

$$\mathbb{P}(A and B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C)$$

Equivalent definition holds for random variables.

We will write $X \perp Y | Z$.

Chain Rule

Definition

For a given set of events $\{S_i\}$:

$p(S_1 \text{ and } S_2 \text{ and } \dots \text{ and } S_n) = p(S_1)p(S_2|S_1)\dots p(S_n|S_1 \text{ and } \dots \text{ and } S_{n1})$

Note that the amount of parameters remain the same:

$p(x_2|x_1 = 0)$ and $p(x_2|x_1 = 1)$ are parameterised by two parameters.

Structure through Chain Rule

- Using Chain Rule

$$p(x_1, \ldots, x_n) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \cdots p(x_n \mid x_1, \cdots, x_{n-1})$$

- How many parameters? $1 + 2 + \cdots + 2^{n-1} = 2^n - 1$
    - $p(x_1)$ requires 1 parameter
    - $p(x_2 \mid x_1 = 0)$ requires 1 parameter, $p(x_2 \mid x_1 = 1)$ requires 1 parameter
      Total 2 parameters.
    - $\cdots$
- $2^n - 1$ is still exponential, chain rule does not buy us anything.
- Now suppose $X_{i+1} \perp X_1, \ldots, X_{i-1} \mid X_i$, then

$$\begin{aligned} p(x_1, \ldots, x_n) &= p(x_1)p(x_2 \mid x_1)p(x_3 \mid \cancel{x_1}, x_2) \cdots p(x_n \mid \cancel{x_1, \cdots,} x_{n-1}) \\ &= p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2) \cdots p(x_n \mid x_{n-1}) \end{aligned}$$

- How many parameters? $2n - 1$. Exponential reduction!

Estimation

Likelihood

Notice, that when we write PDF, we did not assume anything about parameters. What if know the data:

$$\mathbb{P}(data|parameters)\Big|_{dataobs.} = \mathcal{L}(parameters)$$

$\mathcal{L}$ is called the Likelihood Function.
NB: it's not a probability.

Maximum Likelihood Estimator

Definition

Maximum Likelihood Estimator (MLE) is defined as the estimate $\widehat{\theta}_n$ of parameter θ, which maximizes likelihood: $\mathcal{L}_n(\theta)$ (with n being the number of events in a sample).

Some MLE properties

1. MLE is consistent: $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta$.
2. MLE does not depend on the parameterisation: $\widehat{\theta}_n$ — MLE for θ, than $g(\widehat{\theta}_n)$ — MLE for $g(\theta)$;
3. MLE is asymptotically normal: $(\widehat{\theta} - \theta_*)/\widehat{se} \rightsquigarrow \mathcal{N}(0, 1)$;
4. MLE is asymptotically optimal.

Example of MLE:

Find $\widehat{\mu}$ and $\widehat{\sigma}$ for Normal function with number of events in sample n:

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Rewrite as log-likelihood:

$$\ell_n(\mu,\sigma) = \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi}} - \ln\sigma - \frac{(x-\mu)^2}{2\sigma^2} \right)$$

Take derivatives:

$$\frac{\partial \ell_n}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} \quad \frac{\partial \ell_n}{\partial \sigma} = \sum_{i=1}^{n} \left(\frac{(x_i-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \right)$$

Example of MLE:

Thus:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

and:

$$\widehat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2$$

MLE estimate gives not biased $\hat{\sigma}$!

Generative Modeling

Quote

All models are generative models.

-Eric Jang

Generative vs Discriminative Modeling

Discriminative model

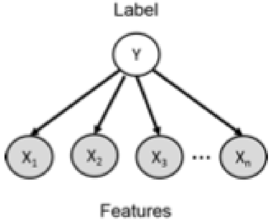
- learn $\mathbb{P}(y|x)$
- Directly characterizes the decision boundary between classes only
- Examples: Logistic Regression, SVM, etc

Generative model

- learn $\mathbb{P}(x|y)$ (and eventually $\mathbb{P}(y,x)$)
- Characterize how data is generated (distribution of individual class)
- Examples: Naive Bayes, HMM, etc.

Naive Bayes

- 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 $p(Y, X_1, \ldots, X_n)$
- Words are conditionally independent given Y:



- Then

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

Naive Bayes: Discrimination

- 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 $p(Y, X_1, \ldots, X_n)$
- Suppose that the words are conditionally independent given Y. Then,

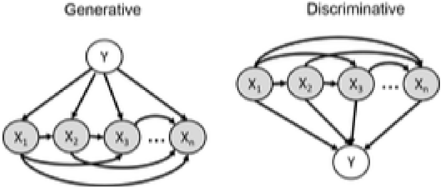
$$p(y, x_1, \ldots x_n) = p(y) \prod_{i=1}^{n} p(x_i \mid y)$$

Estimate parameters from training data. **Predict** with Bayes rule:

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

Discriminative vs Generative Modeling

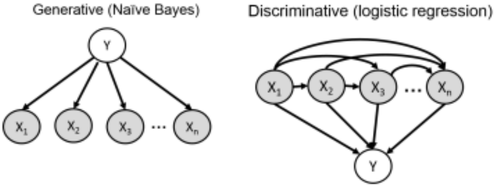
- Since $\mathbf{X}$ is a random vector, chain rules will give
  - $p(Y, \mathbf{X}) = p(Y)p(X_1 \mid Y)p(X_2 \mid Y, X_1) \cdots p(X_n \mid Y, X_1, \cdots, X_{n-1})$
  - $p(Y, \mathbf{X}) = p(X_1)p(X_2 \mid X_1)p(X_3 \mid X_1, X_2) \cdots p(Y \mid X_1, \cdots, X_{n-1}, X_n)$



We must make the following choices:

1. In the generative model, $p(Y)$ is simple, but how do we parameterize $p(X_i \mid \mathbf{X}_{pa(i)}, Y)$?
2. In the discriminative model, how do we parameterize $p(Y \mid \mathbf{X})$? Here we assume we don't care about modeling $p(\mathbf{X})$ because $\mathbf{X}$ is always given to us in a classification problem

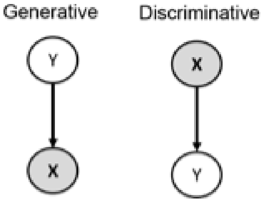
Discriminative outcome



- Logistic model does *not* assume $X_i \perp \mathbf{X}_{-i} \mid Y$, unlike naive Bayes
- This can make a big difference in many applications
- For example, in spam classification, let $X_1 = 1[\text{“bank” in e-mail}]$ and $X_2 = 1[\text{“account” in e-mail}]$
- Regardless of whether spam, these always appear together, i.e. $X_1 = X_2$
- Learning in naive Bayes results in $p(X_1 \mid Y) = p(X_2 \mid Y)$. Thus, naive Bayes **double counts the evidence**
- Learning with logistic regression sets $\alpha_1 = 0$ or $\alpha_2 = 0$, in effect ignoring it

Generative outcome

Using chain rule $p(Y, \mathbf{X}) = p(\mathbf{X} \mid Y)p(Y) = p(Y \mid \mathbf{X})p(\mathbf{X})$. Corresponding Bayesian networks:



1. Using a conditional model is only possible when $\mathbf{X}$ is always observed
   - When some X_i variables are unobserved, the generative model allows us to compute $p(Y \mid \mathbf{X}_{evidence})$ by marginalizing over the unseen variables

Testing the outcome

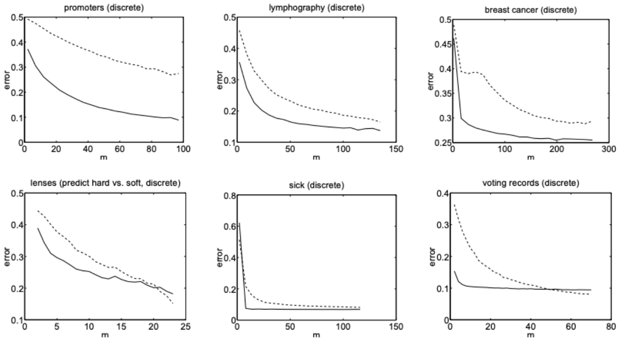


Figure 1: Results of 15 experiments on datasets from the UCI Machine Learning repository. Plots are of generalization error vs. m (averaged over 1000 random train/test splits). Dashed line is logistic regression; solid line is naive Bayes.

Indeed in case we have not enough events, Naive Bayes tend to win.

From A. Ng et al.

Generative Modeling: problem statement

Three major tasks, given a generative model f from a class of models $\mathcal{F}$:

1. Estimation: find the f in $\mathcal{F}$ that best matches observed data.
2. Evaluate Likelihood: compute $f(z)$ for a given z.
3. Sampling: drawing from f.

From S. Nowozin et al.

Sampling ideas

If we have a parametric model, the life simplifies dramatically:

- Specify a latent $p(z)$ followed by a procedure $f_\theta : Z \to X$.
- Key point: in this setting, sampling data is almost always easy.
- Sometimes the whole problem is easy: remember inversion sampling?

  $$z \sim Unif(0;1); x = F_\phi^{-1}(z); x \sim Exp(\phi).$$

  Here F_ϕ is the CDF of the exponential distribution, $F_\phi(x) = 1 - \exp(x\phi)$, with $F_\phi^{-1}(z) = -\phi \log(1 - z)$.
- Unfortunately, more often f_θ induces an intractable log likelihood.

Taxonomy of Generative Model Techniques

- Nonparametric
  - histograms
  - kernel density estimation
- likelihood-based parametric
  - autoregressive models
  - variational autoencoders
  - normalizing flow models
- likelihood-free parametric
  - Generative Adversarial Networks

Wrap up

- Generative modeling includes estimation, evaluation and sampling.
- Some generative models can have problems with components.
- Next: evaluation of Generative models.