Bayesian Methods for Machine Learning

Andrey de Aguiar Salvi

December 2020

1 Class 1

Three principles:

- use prior knowledge
- choose the answer that explains the observations the most
- avoid making extra assumptions

Variable Independence

$$P(X,Y) = P(X)P(Y) \tag{1}$$

Conditional Probability

$$P(X|Y) = \frac{P(X,Y)}{Y(P)} \tag{2}$$

where $P(X,Y)$ = joint probability and $P(Y)$ = marginal probability.

Chain Rule

$$P(X,Y) = P(X|Y)P(Y) \tag{3}$$

$$P(X,Y,Z) = P(X|Y,Z)P(Y|Z)P(Z) \tag{4}$$

$$P(X_1,...,X_N) = \prod_{n=1}^{N} P(X_i|X1,...,X_{i-1}) \tag{5}$$

Marginalization

$$p(X) = \int_{-\inf}^{in} p(X,Y)dY \tag{6}$$

Bayes Theorem

$$P(\theta|X) = \frac{P(X,\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{7}$$

where $P(\theta|X)$ = posterior probability, $P(X|\theta)P(\theta)$ = Likelihood, and $P(X)$ = Evidence.

2 Class 2

Statistic Approaches

- **Frequentist**:
  - deterministic
  - θ is fixed, X is random
  - work whether data points is higher than the parameters - $|X| >> |\theta|$
  - Train models with Maximum Likelihood:

$$\hat{\theta} = arg\max_{\theta} P(X|\theta) \tag{8}$$

    to maximize the probability of data given the parameters

- **Bayesing**:
  - subjective
  - θ is random, X is fixed (given a set of forces θ, tossing a coin always gives the same result X)
  - work with data on any size - $|X|$
  - Train models with Na"(i)ve Bayes to maximize the probability of the parameters given the data

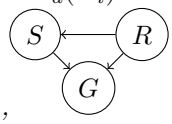
On-line learning: use the current mini-batch (posterior) to update parameters, and then use it as prior in the new mini-batch.

3 Class 3

Bayesian Net - is not a bayesian neural network. The **Nodes** are random variables and the **Edges** are the direct impact. **Model**: is the joint probability over all probabilites

$$P(X_1, ..., X_N) = \prod_{n=1}^{N} P(X_i|P_a(X_i)) \tag{9}$$

where $P_a(X_i)$ is the probability of the parent nodes from the Bayesian Net.

E.g., 

where R is father from G and S, S is parent from G, and the equation bellow is the respective bayesian probability

$$P(S, R, G) = P(G|S, R)P(S|R)P(R) \tag{10}$$

4 Class 5

Univariate Normal Distribution

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{11}$$

Multivariate Normal Distribution

$$\mathcal{N}(x|\mu,\Sigma^2) = \frac{1}{\sqrt{2\pi\Sigma^2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^-1(x-\mu)} \tag{12}$$

Linear Regression

$$L(w) = \sum_{i=1}^{N}(w^T x_i - y_i)^2 = ||w^T X - y|| \rightarrow \min_w \tag{13}$$

$$\hat{w} = arg\min_w L(w) \tag{14}$$

where L is the loss from the bayesian net, w are the weights and X are the data, both are parents of y (target)

$$P(w,y|X) = P(y|X,w)P(w) \tag{15}$$

$$P(y|w,X) = \mathcal{N}(y|w^T X, \sigma^2\mathcal{I}) \tag{16}$$

$$P(w) = \mathcal{N}(w|0,\gamma^2\mathcal{I}) \tag{17}$$

$$P(w|y,x) = \frac{P(y,w|X)}{P(y|X)} \rightarrow \max_w \tag{18}$$

$$P(y,w|x) = \frac{P(y|x,w)}{P(w)} \rightarrow \max_w \tag{19}$$

5 Class 6

Maximum a Posteriori to learn a distribution

$$\theta_{MP} = argmax_\theta P(\theta|X) = argmax_\theta \frac{P(\theta|X)P(\theta)}{P(X)} = argmax_\theta P(\theta|X)P(\theta) \tag{20}$$

We eliminate the denominator from the last equation, once time it don't have θ Problems:

- it is not variant to reparametrization. Ex learning about a gaussian will be not usefull in sigmoid(gaussian)
- strange "loss function"
- we do not have the posteriori of θ
- can't compute credible intervals

6 Class 7

Conjugate Distribution is a way to avoid computing the evidence ($P(X)$), which is costly. In the Bayes Probability

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{21}$$

in the likelihood, $P(X|\theta)$ is fixed by the model, $P(X)$ is fixed by the data, and $P(\theta)$ is our own choice. The prior $P(\theta)$ is conjugate to the likelihood if the prior and the posterior $P(X|\theta)$ lie in the same family distributions.

*E.g.,*if the prior is $P(X|\theta) = \mathcal{N}(x|\mu, \sigma^2)$ and the posterior is $P(\theta) = \mathcal{N}(\theta|m, s^2)$, thus the conjugate of posterior $P(\theta|X)$ is $\mathcal{N}(\theta|a, b^2)$.

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{\mathcal{N}(x|\theta, 1)\mathcal{N}(\theta|0, 1)}{P(X)} \tag{22}$$

$$P(\theta|X) \propto e^{-\frac{1}{2}(X-\theta)^2} e^{-\frac{1}{2}\theta^2} \tag{23}$$

$$P(\theta|X) \propto e^{-(\theta-\frac{X}{2})^2} \tag{24}$$

$$P(\theta|X) = \mathcal{N}(\theta|\frac{X}{2}, \frac{X}{2}) \tag{25}$$

7 Class 8

Gamma Distribution

$$\Gamma(\gamma|a, b) = \frac{b^a}{\Gamma(a)}\gamma^{a-1}e^{-b\gamma} \tag{26}$$

where

- $\gamma, a, b > 0$
- $\Gamma(n) = (n-1)!$
- the expectation, or mean, $\mathbb{E}[\gamma] = \frac{a}{b}$
- $Mode[\gamma] = \frac{a-1}{b}$
- $Var[\gamma] = \frac{a}{b^2}$

Precision

$$\gamma = \frac{1}{\sigma^2} \tag{27}$$

If we replace the variance in the Normal Distribution to the inverse of the Precision, we get

$$\mathcal{N}(x|\mu, \gamma^{-1}) = \frac{\sqrt{\gamma}}{\sqrt{2\pi}} e^{-\gamma\frac{(x-\mu)^2}{2}} \tag{28}$$

thus, the conjugate prior in respect to the precision is

$$\mathcal{N}(x|\mu,\gamma^{-1}) \propto \gamma^{\frac{1}{2}} e^{-b\gamma} \tag{29}$$

$$P(\gamma) \propto \gamma^{a-1} e^{-b\gamma} \tag{30}$$

$$P(\gamma|X) = \Gamma(\gamma|a,b) \tag{31}$$

then, the prior is

$$P(\gamma|X) = \Gamma(\gamma|a,b) \propto \gamma^{a-1} e^{-b\gamma} \tag{32}$$

the posterior is

$$P(\gamma|X) \propto P(X|\gamma)P(\gamma) \tag{33}$$

dropping out all the constants, we have

$$P(\gamma|X) \propto \left(\gamma^{\frac{1}{2}} e^{-\gamma\frac{(X-\theta)^2}{2}}\right)\left(\gamma^{a-1} e^{-b\gamma}\right) \tag{34}$$

re-arranging the terms, we get

$$P(\gamma|X) \propto \gamma^{\frac{1}{2}+a-1} e^{-\gamma\frac{(X-\theta)^2}{2}} e^{-\gamma(b+\frac{(X-\mu)^2}{2})} \tag{35}$$

finally

$$P(\gamma|X) = \Gamma(a+\frac{1}{2}, b+\frac{(X-\mu)^2}{2}) \tag{36}$$

8 Class 9

Beta Distribution

$$\beta(X|a,b) = \frac{1}{\beta(a,b)} X^{a-1}(1-X)^{b-1} \tag{37}$$

where:

- $X \in [0,1]$
- $a, b > 0$
- $$\beta(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \tag{38}$$
- $$\mathbb{E}[X] = \frac{a}{a+b} \tag{39}$$
- $$Mode[X] = \frac{a-1}{a+b-2} \tag{40}$$
- $$Var[X] = \frac{ab}{(a+b)^2(a+b-1)} \tag{41}$$

*E.g.,*to model a distribution with mean 0.8 and standard deviation of 0.1, thus $\mathbb{E}[X] = 0.8$ and $Var[X] = 0.1^2$. Consequently, a Beta distribution with $a = 12$ and $b = 3$ model it.

Bernoulli Distribution The Beta distribution is the conjugate of the Bernoulli likelihood. *E.g.,*a Bernoulli likelihood from a dataset

$$P(X|\theta) = \theta^{N_1}(1-\theta)^{N_0} \tag{42}$$

where N_1 is the number of ones in X and N_0 the number of zeros. Thus, the Beta distribution is

$$P(\theta) = \beta(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1} \tag{43}$$

Multiplying the likelihood by the prior

$$P(\theta|X) \propto P(X|\theta)P(\theta) \tag{44}$$

using the before, but replacing the terms by the equivalent equations above, we have

$$P(\theta|X) \propto \theta^{N_1}(1-\theta)^{N_0}\beta(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1} \tag{45}$$

and rearranging the terms, we have

$$P(\theta|X) \propto \theta^{N_1+a-1}(1-\theta)^{N_0+b-1} \tag{46}$$

and finally, recognizing the equation before as a Beta distribution, we have

$$P(\theta|X) = \beta(N_1+a, N_0+b) \tag{47}$$

Posteriors

- Pros
  - Exact posterior
  - Easy for on-line learning. *E.g.,*$P(\theta|X) = \beta(N_1+a, N_2+b)$
- Cons
  - In some cases, the conjugate prior may be inadequate

9 Class 10

Latent Variable is a hidden variable that you never observe. *E.g.,*creating a dataset of a job interview, measuring the GPA, IQ, School degree, and Phone, from a first stage of the interview, and a last attribute Onsite performance, from the second interview. Traditional ML models will suffer with the missing data. Even more, probably there are some correlation between the variables, and ponder each combination will increase the number of attributes. Thus, we

can link all these attributes with a latent variable, called in this example as Intelligence. We can model the problem as:

$$P(X_1, X_2, X_3, X_4, X_5) = \sum_{i=1}^{N} P(X_1, X_2, X_3, X_4, X_5|I)P(I) \tag{48}$$

where I is the latent variable. We can also simplify the problem with

$$P(X_1, X_2, X_3, X_4, X_5) = \sum_{i=1}^{N} P(X_1|I)P(X_2|I)P(X_3|I)P(X_4|I)P(X_5|I)P(I) \tag{49}$$

which breaks the table in 5 fewer tables, reduce the model complexity and improves the flexibility of the model.

10 Class 12

Gaussian Mixture Model (GMM) A model of soft clustering with N gaussian's can be described as

$$P(X|\theta) = \pi_1\mathcal{N}(\mu_1, \Sigma_1) + \pi_2\mathcal{N}(\mu_2, \Sigma_2) + ... + \pi_N\mathcal{N}(\mu_N, \Sigma_N) \tag{50}$$

where the θ weight matrix is

$$\theta = \pi_1, \pi_2, ..., \pi_N, \mu_1, \mu_2, ..., \mu_N, \Sigma_1, \Sigma_2, ..., \Sigma_N \tag{51}$$

To goal of the train is

$$\max_{\theta} P(X|\theta) = \prod_{i=1}^{N} P(X_1|\theta) = \prod_{i=1}^{N} (pi_1\mathcal{N}(\mu_1, \Sigma_1) + ...) \tag{52}$$

By definition, the covariance matrix $\Sigma_k \succ 0$, once time we need to compute Σ^{-1}. Otherwise, we will have divisions by zero.

11 Class 13

Assuming that our data points X was generated by a latent variable t. Thus, the probability of a point to belongs to the class c given the parameters θ is

$$P(t = c|\theta) = \pi_c \tag{53}$$

and

$$P(X|t = c, \theta) = \mathcal{N}(X, \mu_c, \Sigma_c) \tag{54}$$

Marginalizing the latent variable, we have

$$P(X|\theta) = \sum_{c=1}^{N} P(X|t = c, \theta)P(t = c|\theta) \tag{55}$$

which is the summation of the likelihood times the prior, and is exactly the same as the first equation, but ignoring the latent variable.

Training a GMM, if we have hard-labels of the points, it is a easy problem: we just need to compute the mean and standard deviation of each cluster. In the opposite side, if we have the parameters (consequently the gaussians), thus we can estimate the source ($P(t = 1|X, \theta) = \frac{P(X|t=1,\theta)P(t=1|\theta)}{Z}$), which is the joint probability (the likelihood times the prior). The problem of train a GMM is a Chicken and Egg problem:

- Need gaussian parameters (θ) to estimate the source (soft-labels)
- Need sources to estimate the gaussian parameters