ANLY601 Take Home Assignment 2

Xunge Jiang

Jan 27, 2019

1 Assume a Single Gaussian Classifier with training data X. Assume a fixed mean parameter μ. What is the Maximum Likelihood Objective one would use to solve for the variance parameter σ?

$$\mathcal{L}(\sigma) = \arg\max_{\theta} P(X|\sigma) = \arg\max_{\theta} \prod_{i=1}^{N} P(x_i|\sigma)$$

Or using the log likelihood which is

$$\mathcal{L}(\sigma) = \arg\max_{\theta} lnP(X|\sigma) = \arg\max_{\theta} \prod_{i=1}^{N} lnP(x_i|\sigma) = \arg\max_{\theta} \sum_{i=1}^{N} lnP(x_i|\sigma)$$

2 Find (solve for) the ML estimate for the variance. Show all steps starting with the objective in #1. Show / explain all steps

We know that we can write Gaussian (where σ represents variance here instead of standard deviation) as:

$$ln\mathcal{L}(\sigma) = \textstyle\sum_{i=1}^{N} lnP(x_i|\sigma) = \sum_{i=1}^{N} ln\frac{1}{2\pi^{1/2}} + ln\frac{1}{\sigma^{1/2}} - \frac{1}{2}(x_i-u)\sigma^{-1}(x_i-u)^T$$

Since first term is not related to σ, therefore,

$$\frac{\partial\mathcal{L}(\sigma)}{\partial\sigma} = \sum_{i=1}^{N} -\frac{1}{2}\sigma^{\frac{1}{2}}\sigma^{-\frac{3}{2}} + \frac{1}{2}(x_i-u)\sigma^{-2}(x_i-u)^T$$

$$0 = \sum_{i=1}^{N} -\frac{1}{2\sigma} + \frac{(x_i-u)(x_i-u)^T}{2\sigma^2}$$

$$N\sigma = \sum_{i=1}^{N}(x_i - u)(x_i - u)^T$$

$\Rightarrow$ Maximum likelihood of variance using parameter σ:

$$\boxed{\sigma_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - u)(x_i - u)^T}$$

3 Show the MAP objective and assuming a fixed mean μ. Show all steps starting with the MAP objective. Show / explain all steps

MAP Objective:

$$\mathcal{L}(\mu) = \arg\max_{\theta} P(X|\mu)P(\mu) = \arg\max_{\theta} \prod_{i=1}^{N} P(x_i|\mu)P(\mu)$$

Or using the log likelihood which is

$$\mathcal{L}(\mu) = \arg\max_{\theta} ln(P(X|\mu)P(\mu)) = \arg\max_{\theta} \prod_{i=1}^{N} ln(P(x_i|\mu)P(\mu)) = \arg\max_{\theta} \sum_{i=1}^{N} lnP(x_i|\mu) + lnP(\mu)$$

Denote Σ to be variance for likelihood $P(x_i|\mu)$ and Σ_0 to be variance for prior $P(\mu|\mu_0, \Sigma_0)$. Then for MAP,

$P(x_i|\mu) = \frac{1}{2\pi^{1/2}(\Sigma)^{1/2}} e^{-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}$

$P(\mu|\mu_0, \Sigma_0) = \frac{1}{2\pi^{1/2}(\Sigma_0)^{1/2}} e^{-\frac{1}{2}(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)}$

$\Rightarrow ln\mathcal{L}(\mu) = \sum_{i=1}^{N} lnP(x_i|\mu) + lnP(\mu)$

$$= \sum_{i=1}^{N} ln(\frac{1}{2\pi^{1/2}\Sigma^{1/2}}) - \frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) + ln(\frac{1}{2\pi^{1/2}\Sigma_0^{1/2}}) - \frac{1}{2}(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)$$

Since first term and third term are not related to μ, thus,

$$\frac{\partial\mathcal{L}(\mu)}{\partial\mu} = \sum_{i=1}^{N}\Sigma^{-1}(x_i - \mu) - \Sigma_0^{-1}(\mu - \mu_0) = 0$$

$$\Sigma^{-1}\sum_{i=1}^{N} x_i - \Sigma^{-1}N\mu = \Sigma_0^{-1}(\mu - \mu_0) = 0$$

$$\Sigma^{-1}\sum_{i=1}^{N} x_i + \Sigma_0^{-1}\mu_0 = \Sigma_0^{-1}\mu + \Sigma^{-1}N\mu$$

$$\frac{\Sigma_0 \sum_{i=1}^{N} x_i + \Sigma\mu_0}{\Sigma\Sigma_0} = \frac{(\Sigma + \Sigma_0 N)\mu}{\Sigma\Sigma_0}$$

Times both sides by $\Sigma\Sigma_0$, therefore, we get

$$\mu = \frac{\Sigma_0 \sum_{i=1}^{N} x_i + \Sigma\mu_0}{\Sigma + \Sigma_0 N}$$

Take $\theta = \sqrt{\Sigma}$ and $\theta_0 = \sqrt{\Sigma_0}$, we can achieve the same equation given by Dr. Bolton,

$$\boxed{\hat{\mu}_{MAP} = \frac{\frac{\theta_0^2}{\theta^2} \sum_{i=1}^{N} x_i + \mu_0}{1 + \frac{\theta_0^2}{\theta^2} N}}$$

4 Briefly compare and contrast the ML and MAP estimates for a single Gaussian Classifier.

ML and MAP both are estimating a single estimation from certain distribution or models.

For MLE:

$$\theta_{MLE} = \arg\max_{\theta} P(X|\theta) = \arg\max_{\theta} \prod_{i} P(x_i|\theta)$$

Or Using Log:

$$\theta_{MLE} = \arg\max_{\theta} \log P(X|\theta) = \arg\max_{\theta} \log \prod_{i} P(x_i|\theta) = \arg\max_{\theta} \sum_{i} \log P(x_i|\theta)$$

For MAP:

It usually comes up in Bayesian setting that it works on a posterior distribution, not only the likelihood (one more component).

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

$$\theta_{MAP} = \arg\max_{\theta} P(X|\theta)P(\theta)$$

Thus, two methods are differed by inclusion of prior in MAP.

Then:

- Case 1: If we have uniform prior ($P(\theta)$ constant everywhere in the distribution), MAP estimation can be ignored. Two estimations are very similar.

- Case 2: If N is sufficiently large, then MAP and MLE estimation are very similar. Variance of two methods have the same equation, only compare the estimation of μ.

  Proof:

$$\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\mu_{MAP} = \frac{\Sigma_0 \sum_{i=1}^{N} x_i + \Sigma\mu_0}{\Sigma + \Sigma_0 N}, N \to \infty, \mu_{MAP} = \frac{\Sigma_0 \sum_{i=1}^{N} x_i}{\Sigma_0 N} = \frac{1}{N}\sum_{i=1}^{N} x_i = \mu_{MLE}$$

- Case 3: If variance (Σ_0) of prior is sufficiently large, then MAP and MLE estimation are also very similar. Again, variance of two methods have the same equation, only compare the estimation of μ.

  Proof:

$$\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\mu_{MAP} = \frac{\Sigma_0 \sum_{i=1}^{N} x_i + \Sigma\mu_0}{\Sigma + \Sigma_0 N}, \Sigma_0 \to \infty, \mu_{MAP} = \frac{\Sigma_0 \sum_{i=1}^{N} x_i}{\Sigma_0 N} = \frac{1}{N}\sum_{i=1}^{N} x_i = \mu_{MLE}$$

$\boxed{\text{Conclusion:}}$

- 1. They are differed by the information of prior (if prior is very informative.)

- 2. By three cases above, they are very similar, which we can say that MLE is a special case of MAP.

5 Prove that the ML and MAP mean estimates are similar when the prior probability is uniform. Prove this semi-formally – use math to help support your claims. ALSO Create plots from your in-class exercise to support your claim.

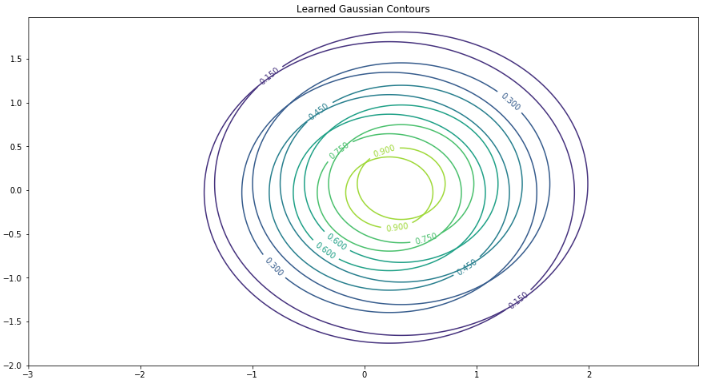
$$\hat{\theta}_{MAP} = \arg\max_{\theta} \sum_{i} \log P(x_i|\theta)P(\theta) \tag{1}$$

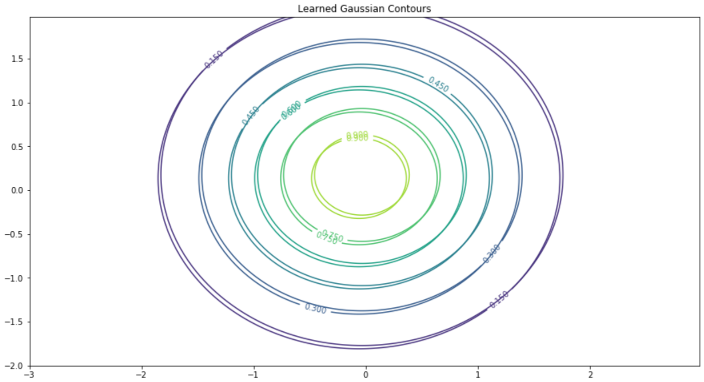
$$= \arg\max_{\theta} \sum_{i} \log P(x_i|\theta)\, const \tag{2}$$

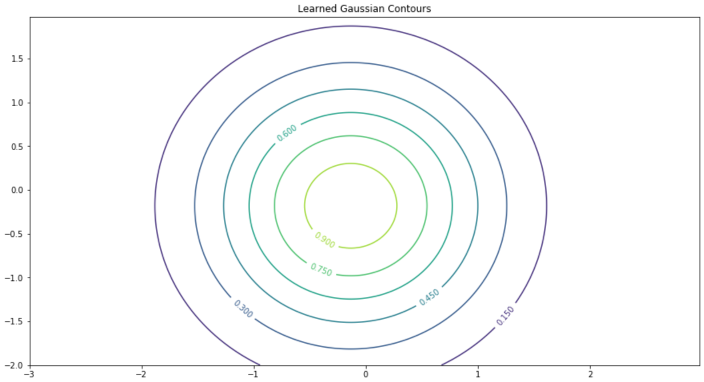
$$\approx \arg\max_{\theta} \sum_{i} \log P(x_i|\theta) \tag{3}$$

$$= \hat{\theta}_{MLE}$$

Due to the random generation of uniform distribution, three plots were generated when prior is uniform to ensure the finding is true. As we can see from three graphs, ML and MAP mean estimates are very similar as they overlap/close to overlap.




Learned Gaussian Contours
0.150
0.300
0.450
0.600
0.750
0.900
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−3
−2
−1
0
1
2


Learned Gaussian Contours
0.150
0.300
0.450
0.600
0.750
0.900
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−3
−2
−1
0
1
2