CS 5/4780

Probabilistic modeling, MLE and MAP Estimates

Recall the ML Setup: $(X, Y) \sim P$

If we knew $P(X, Y)$ or even just $P(Y|X)$, we could compute Bayes optimal classifier

For classification

$$h(x) = \operatorname*{argmax}_{y \in C} P(Y = y \mid X = x)$$

more generally

$$h(x) = \operatorname*{argmin}_{\hat{y}} E_{y|x}[\ell(\hat{y}, y)]$$

Generative: $P(X|Y)\,P(Y)$

Probabilistic modeling: Estimate $P(X,y)$

Discriminative (or $P(Y|X)$ directly) and use it instead

Estimating Bernoilli R.V.: Yearly rain/no Rain Data

$$D = \{R, N, N, N, R, N, N\}$$

R = Rain N = No Rain

$\mathcal{X} = \{\}$

$C = \{R, N\}$

What would your estimate to $P(Y)$ be?

$$\boxed{}$$

Can we derive this formally?

Assume events are "Independent and Identically distributed" i.i.d.

Parameter: $p = P(Y = R)$ n_R = # Rainy days

n_N = # no rain days

$$\hat{p} = \underset{p\in[0,1]}{\operatorname{argmax}} \binom{n_R+n_N}{n_R} p^{n_R}(1-p)^{n_N} \quad (1)$$

$$= \underset{p\in[0,1]}{\operatorname{argmax}} \log\left(\binom{n_R+n_N}{n_R}\right) + n_R \log p + n_N \log(1-p) \quad (2)$$

$$= \underset{p\in[0,1]}{\operatorname{argmax}} \; n_R \log p + n_N \log(1-p) \quad (3)$$

$$(4) \Rightarrow \frac{n_R}{\hat{p}} - \frac{n_N}{1-\hat{p}} = 0 \Rightarrow \hat{p} = \frac{n_R}{n_R+n_N}$$

1. Parameterize $P(\text{Data})$ by some family of parameters P_θ s.t. $\theta \in \Theta$
2. Estimate $P(x,y)$ (or $P(y|x)$) by estimating $\hat{\theta} \in \Theta$ from Data

MLE : Maximum Likelihood Estimate

$$\theta_{MLE} = \underset{\theta\in\Theta}{\operatorname{argmax}} \; P_\theta(\text{Data})$$

pick θ that maximizes likelihood of Data observation

1. Often referred to as frequentist view
2. when $\theta^* \in \Theta$ generates data, $\theta_{MLE} \to \theta^*$ (typically)

Eg 2: Data : $D = \{176, 177, 169, 168, \ldots\}$
Heights of Adult Male / Female

heights are normally distributed with mean μ and variance σ^2

$$\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \; \sigma^2 \in \mathbb{R}^+\}$$

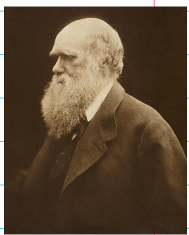
① $$\left(\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE}\right) = \underset{\mu, \sigma^2}{\arg\max} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right)$$

② $$\left(\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE}\right) = \underset{\mu, \sigma^2}{\arg\max} \; -\sum_{i=1}^{n}\left(\frac{(x_i-\mu)^2}{2\sigma^2} + \log\left(\frac{1}{\sigma}\right)\right)$$

③ $$\frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_{MLE})}{\sigma^2} = 0, \qquad \sum_{i=1}^{n}\left(\frac{(x_i-\hat{\mu}_{MLE})^2}{\hat{\sigma}^3_{MLE}} - \frac{1}{\hat{\sigma}_{MLE}}\right) = 0$$

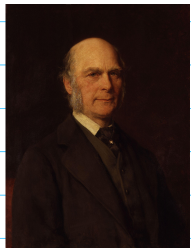
④ $$\left(\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE}\right) = \left(\underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}_{\text{Sample mean}}, \underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{MLE})^2}_{\text{Sample variance}}\right)$$

Gaussian Mixture Model :



Charles Darwin

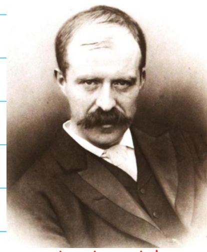
Evolution via Natural selection



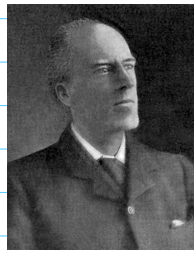
Francis Galton

Evolution discontinuous sudden

Vs

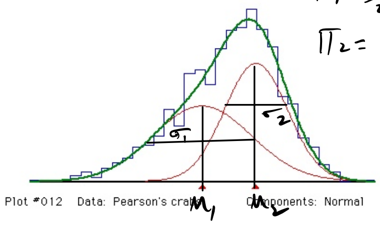


Raphael Weldon



Karl Pearson

small variations, gradual evolution



$\pi_1 = \frac{1}{2}$

$\pi_2 = \frac{1}{2}$

$$\Theta = \begin{cases} \pi = (\pi_1, \ldots, \pi_K) & \text{distribution over } K \text{ items} \\ \mu_1, \ldots, \mu_K \in \mathbb{R}^d & K\text{- means} \\ \Sigma_1, \ldots, \Sigma_K \in \mathbb{R}^{d \times d} & K\text{- covariances} \end{cases}$$

Normal pdf

$$\hat{\theta}_{MLE} = \underset{\pi, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K}{\arg\max} \quad P_\theta(D) := \prod_{i=1}^{n}\left(\sum_{k=1}^{K} \pi_k \, p(x_i \mid \mu_k, \Sigma_k)\right)$$

EM Algo for GMM aims to solve above

MLE does not capture prior knowledge

Eg Rain, No Rain.

Say we had prior info that at similar locations typically we have seen Rain on 30 out of 100 days, how do we use this?

Heuristic : $p = P(Y=\text{Rain}) = \dfrac{n_R + 30}{n_R + n_N + 100}$

Maximum Aposteriori Estimator : MAP

Model is an abstraction that captures our belief, we update our belief based on Data.

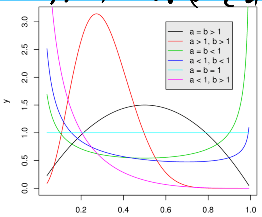
θ is a Random variable

$$\hat{\theta}_{MAP} = \underset{\theta \in \Theta}{\operatorname{argmax}} \; P(\theta \mid D) = \underset{\theta \in \Theta}{\operatorname{argmax}} \; P(D \mid \theta) P(\theta)$$

$$= \underset{\theta \in \Theta}{\operatorname{argmax}} \; \underbrace{\log(P(D \mid \theta))}_{\text{log likelihood}} + \underbrace{\log P(\theta)}_{\text{log prior}}$$

Eg: For Bernoulli distribution we can use Beta prior

$$P(\theta) = \frac{\theta^{\alpha - 1} (1-\theta)^{\beta - 1}}{B(\alpha, \beta)}$$



(1) $\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \; \log P(D \mid \theta) + \log P(\theta)$

(2) $= \underset{\theta}{\operatorname{argmax}} \; n_R \log\theta + n_N \log(1-\theta) + (\alpha - 1)\log\theta + (\beta - 1)\log(1-\theta) - \log B(\alpha, \beta)$

(3) $\hat{\theta}_{MAP} = \dfrac{n_R + \alpha - 1}{n_R + n_R + \alpha + \beta - 2}$

$\alpha - 1$ Rains

$\beta - 1$ no Rains

Often MAP is referred to as Bayesian view

There is Bayesian and there is BAYESIAN

True Bayesian: "There is no model, all you are estimating is Y"

$$P(Y \mid X, Data) = \int_{\theta} P(Y, \theta \mid X, Data)\, d\theta$$

$$= \int_{\theta} P(Y \mid \theta, X, Data)\, P(\theta \mid Data)\, d\theta$$