CS 3780/5780

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:

$$h(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \; P(Y = y \mid X = x)$$

more generally:

$$h(x) = \underset{g}{\operatorname{argmin}} \; E_{Y|X}[\ell(g, y)]$$

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

Probabilistic modeling: Estimate $P(X, y)$ (or $P(Y|X)$ directly) and use it instead

Discriminative

Estimating Bernouilli R.V.: Yearly rain/no Rain

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

R = "Rain", N = "No Rain"

$X = \{\}$

$Y = \{R, N\}$

What would your estimate for $P(Y)$ be given data D?

Can we derive this formally? $\hat{P}(y = \text{rain}) = 2/7$

Modeling assumption: rain/no rain drawn i.i.d.

"Independent and Identically distributed"

$p = P(Y = R)$

n_R = # Rainy days $= 2$

n_N = #no rain days $= 5$

What is the likelihood of data D under model with parameter p?

$$\text{Likelihood}(D) = \binom{n_R + n_N}{n_R} p^{n_R} (1-p)^{n_N}$$

"All models are wrong,
...but some are useful"
- George Box

1. Parameterize $P(X,Y)$ by some family of distributions P_θ s.t. $\theta \in \Theta$

2. Estimate $P(X, Y)$ (or $P(Y|X)$) by picking $\theta \in \Theta$ based on Data D

Maximum Likelihood estimator: Pick $\theta \in \Theta$ that maximizes likelihood of observation of data D

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

$$= \underset{\theta \in \Theta}{\operatorname{argmax}} \; \log P_\theta(D) \quad \text{"log likelihood"}$$

↳ i.i.d data $P_\theta(D) = \prod_{i=1}^{n} P_\theta(x_i, y_i)$

1 Often referred to as frequentist view
2 when $\theta^* \in \Theta$ generates the data, $\hat{\theta}_{MLE} \to \theta^*$ (As $n \to \infty$)

Steps to compute MLE:

1. Pick the probabilistic model, and identify parameters for that model
2. Write down the likelihood of data under model parameter
3. Write down the log likelihood of data D and simplify the expression
4. Maximize the expression w.r.t. parameter and hence find the MLE estimate

Eg 1. Rain Data $D = \{R, N, N, N, R, N, N\}$

① $\theta = p = P(y = R)$ $\quad \Theta = [0, 1]$

② $\text{Likelihood}_p(D) = \binom{n_R + n_N}{n_R} p^{n_R} (1-p)^{n_N}$

③ $\hat{p}_{MLE} = \underset{p \in [0,1]}{\text{argmax}} \ \text{Likelihood}_p(D)$

$= \underset{p \in [0,1]}{\text{argmax}} \ \log(P_\theta(D))$

$\hat{p}_{MLE} = \underset{p}{\text{argmax}} \ \log\left(\binom{n_R + n_N}{n_R} p^{n_R}(1-p)^{n_N}\right)$

$\hat{p}_{MLE} = \underset{p}{\text{argmax}} \ n_R \log(p) + n_N \log(1-p)$

Eg 2: 1. Heights of Adult Male (or female)
2. Shoe size
3. Blood pressure

~~$D = \{176, 177, 169, 168, \ldots\}$~~

$\Theta = (\mathbb{R}, \mathbb{R}^+)$ Normally distributed $\quad D = \{x_1, \ldots, x_n\}$

① $\theta = (\mu, \sigma^2)$ (parameters) $\quad p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

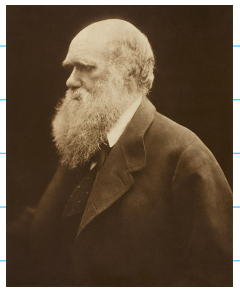
② $P_\theta(D) = \prod_{i=1}^{n} p_\theta(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

③ $\underset{\mu, \sigma^2}{\text{argmax}} \ \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)$

$= \underset{\mu, \sigma^2}{\text{argmax}} \ \sum_{i=1}^{n} \left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right]$

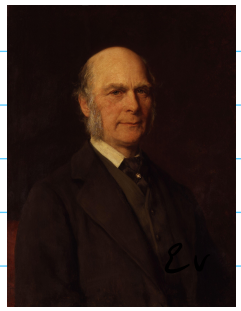
$= \underset{\mu, \sigma^2}{\text{argmax}} \left[-\frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right]$

Gaussian Mixture Model :



Charles Darwin

Evolution via Natural Selection



Francis Galton

Vs

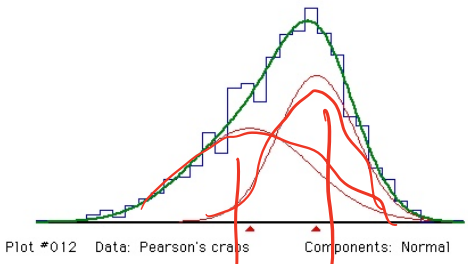
Raphael Weldon

Karl Pearson

Evolution is discontinuous

Evolution is gradual (small changes over generations)

crabs from Naples



(p, μ, σ)

Forehead breadth to body length

p_1, p_2

(1) parameters

$p = (p_1, p_2)$ p_1 = prob of crab 1, p_2 = prob of crab 2

$\mu = (\mu_1, \mu_2)$ mean feature for each of the species

$\sigma^2 = (\sigma^2_1, \sigma^2_2)$ var for each species

MLE does not capture prior knowledge

Eg1. 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 : $p = P(Y = \text{Rain}) = \dfrac{n_R + 2}{n_R + n_N + 365}$

Maximum Aposteriori Estimator : MAP

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

θ is a Random variable

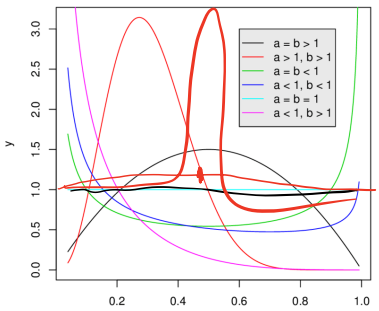
$P(\theta)$ prior

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

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

$P_\theta(D)$

Rain/No Rain Eg. Beta prior:



① $P(\theta) = \dfrac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}$ Prior

$P_\theta(Y = \text{rain}) = \theta$ likelihood

② $\hat{\theta}_{MAP} = \log(P(D \mid \theta)) + \log P(\theta)$

$= n_R \log\theta + n_N \log(1-\theta) + (a-1)\log\theta + (b-1)\log(1-\theta)$

③ taking derivative $= 0$ $\quad \dfrac{n_R}{\hat{\theta}_{MAP}} - \dfrac{n_N}{1-\hat{\theta}_{MAP}} + \dfrac{a-1}{\hat{\theta}_{MAP}} - \dfrac{b-1}{1-\hat{\theta}_{MAP}} = 0$

④ $\hat{\theta}_{MAP} = \dfrac{n_R + a - 1}{n_R + n_N + a + b - 2}$

$a-1$: Rains

$b-1$: 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".

$$P(Y|X,D) = \int_{\theta} P(Y,\theta|X,D)\, d\theta$$

$$= \int_{\theta} P(Y|\theta,X,D)\, P(\theta|D)\, d\theta$$

$$\hat{p}_{MLE} = \underset{p}{\operatorname{argmax}} \log\left[\binom{n_R+n_N}{n_R} p^{n_R}(1-p)^{n_N}\right]$$

$$= \underset{p}{\operatorname{argmax}} \log\left(\binom{n_R+n_N}{n_R}\right) + n_R \log(p) + n_N \log(1-p)$$

To optimise

$$\frac{d}{dp}\left(n_R \log p + n_N \log(1-p)\right) = 0$$

$$\frac{n_R}{p} - \frac{n_N}{1-p} = 0$$

$$\hat{p}_{MLE} = \frac{n_R}{n_R + n_N} = 2/7$$

$$\hat{h}(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \; P_{\hat{\theta}_{MLE}}(Y = y \mid X)$$

$$P(\theta \mid D) = \frac{P(D \mid \theta) P(\theta)}{P(D)}$$

$$\underset{\theta}{\operatorname{argmax}} P(\theta \mid D) = \underset{\theta}{\operatorname{argmax}} \log P(\theta \mid D)$$

$$= \underset{\theta}{\operatorname{argmax}} \log \left(\frac{P(D \mid \theta) P(\theta)}{P(D)} \right)$$

$$= \underset{\theta}{\operatorname{argmax}} \log P(D \mid \theta) + \log P(\theta)$$

log likelihood

log prior

$- \log P(D)$

$$\underset{\mu,\sigma^2}{\operatorname{argmax}} \sum_{i=1}^{n} \left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(x_i-\mu)^2}{2\sigma^2} \right]$$

$$L(\mu,\sigma)$$

$$\frac{\partial L(\mu,\sigma)}{\partial \mu} = 0$$

$$= \sum_{i=1}^{n} \frac{+2(x_i-\mu)}{2\sigma^2} = 0$$

$$= \sum_{i=1}^{n} \frac{(x_i-\mu)}{\sigma^2} = 0$$

$$\sum_{i=1}^{n} (x_i - \mu) = 0$$

$$\sum_{i=1}^{n} x_i - n\mu = 0$$

$$\hat{\mu}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\frac{\partial L(\mu,\sigma^2)}{\partial \sigma^2} = \frac{\partial}{\partial \sigma^2}\left(\sum_{i=1}^{n} -\frac{1}{2}\log 2\pi\sigma^2 - \frac{(x_i-\mu)^2}{2\sigma^2} \right)$$

$$0 = \frac{\partial}{\partial \sigma^2}\left(\sum_{i=1}^{n} -\frac{1}{2}\log\sigma^2 - \frac{(x_i-\mu)^2}{2\sigma^2} \right)$$

$$\theta = \cdot \sum_{i=1}^{n} - \frac{1}{\sigma^2} + \frac{(x_i - \mu)^2}{\sigma^4 2}$$

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

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

$$P(\theta \mid D)$$

$$= \frac{P(D \mid \theta)\ P(\theta)}{P(D)}$$

$$\underset{\theta}{\operatorname{argmax}}\ P(\theta \mid D)$$

$$= \underset{\theta}{\operatorname{argmax}}\ \frac{P(D \mid \theta)\ P(\theta)}{P(D)}$$

$$= \arg\max_{\theta} \log P(D|\theta) P(\theta)$$

$$= \arg\max_{\theta} \log P(D|\theta) + \log P(\theta)$$