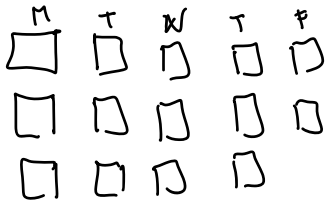
⑦ Statistical Learning

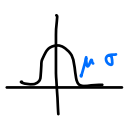


* Introduction

- * Useful for generalization / prediction
- * Concepts: { Parameter estimation, Bias, Variance }
- * Point estimation: { single best prediction of quantity of interest; Example: { vector θ; a function y ← we are interested in estimate the output }
- * { Parameter γ; Point estimate of parameter $\hat{\gamma}$ }
- * Point estimate or statistic $\hat{\gamma}_m = g(x^{(1)}, \dots x^{(m)})$ $\{x^{(i)}\}$ i.i.d data points
- * Good estimator: $\hat{\gamma}_m$ close to γ
- * Frequentist perspective: { γ is fixed but unknown; $\hat{\gamma} = g(x^{(i)})$. Since $x^{(i)}$ is random ⇒ $\hat{\gamma}$ random }
- * Function estimation (= approximation): { $y = f(x) + \varepsilon$; $\hat{f}$ is a point estimator }
  - Example: { Point estimation: $\min \| X^{(test)} \theta - y^{(test)} \|^2$; Function estimation: $\min \| f(X^{(test)}) - y^{(test)} \|$ where $f = X^{(test)}\theta$ }

* Bias

$$\text{bias}(\hat{\gamma}_m) = \mathbb{E}_{x \sim p(x;\gamma)}(\hat{\gamma}_m(x^{(i)})) - \gamma;$$

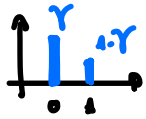


- $\mathbb{E}(\cdot)$ over the data (samples from a random variable)
- γ true underlying value of γ used to define data-generating distribution

* { Unbiased: $\text{bias}(\hat{\gamma}_m) = 0 \Rightarrow \gamma = \mathbb{E}(\hat{\gamma}_m)$; Asymptotically unbiased: $\lim_{m \to \infty} \mathbb{E}(\hat{\gamma}_m) = \gamma$ }

* **Example 1:** Bernoulli Distribution $\{x^{(1)}, \dots x^{(m)}\}$ iid with mean γ

$$P(x^{(i)}, \gamma) = \gamma^{x^{(i)}}(1-\gamma)^{(1-x^{(i)})}$$



How much is γ: $\hat{\gamma}_m = \frac{1}{m}\sum_1^m x^{(i)}$ (empirical mean)

$$\text{bias}(\hat{\gamma}_m) = \mathbb{E}(\hat{\gamma}_m) - \gamma = \mathbb{E}\left(\frac{1}{m}\sum_1^m x^{(i)}\right) - \gamma = \frac{1}{m}\sum_1^m \mathbb{E}(x^{(i)}) - \gamma =$$

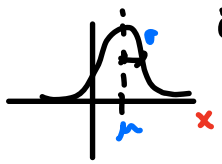
$$= \frac{1}{m}\sum_1^m \sum_{x^{(i)}=0}^{1} P(x^{(i)}|\gamma)\, x^{(i)} - \gamma = \frac{1}{m}\sum_1^m \sum_{x^{(i)}=0}^{1} \gamma^{x^{(i)}}(1-\gamma)^{(1-x^{(i)})} x^{(i)} - \gamma =$$

$$= \frac{1}{m}\sum_1^m \gamma - \gamma = 0$$

* **Example 2:** Gaussian distribution Estimation of the mean

: $\{x^{(1)}, \dots x^{(m)}\}$ iid from $p(x^{(i)}) = N(x^{(i)}; \mu, \sigma)$

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



i σ, μ values ?

How much is μ: $\hat{\mu}_m = \frac{1}{m}\sum_i^m x^{(i)}$

$$\text{bias}(\hat{\mu}_m) = \mathbb{E}(\hat{\mu}_m) - \mu = \mathbb{E}\left(\frac{1}{m}\sum_i^m x^{(i)}\right) - \mu = \frac{1}{m}\sum_i^m \mathbb{E}(x^{(i)}) - \mu$$

$$= \frac{1}{m}\sum_i^m \int x^{(i)} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x^{(i)}-\mu)^2}{\sigma^2}} dx^{(i)} - \mu = \frac{1}{m}\sum_1^m \mu - \mu = 0$$

How much is σ: $\hat{\sigma}_m = \frac{1}{m}\sum_{i=1}^m (x^{(i)} - \hat{\mu}_m)^2$

$$\text{bias}(\hat{\sigma}_m) = \mathbb{E}(\hat{\sigma}_m) - \sigma = \mathbb{E}\left(\frac{1}{m}\sum_{i=1}^m (x^{(i)} - \hat{\mu}_m)^2\right) - \sigma =$$

↑ we will see it later

* **Variance**

$$\text{Var}(\gamma_m) = \mathbb{E}\left[(\gamma_m - \mathbb{E}(\gamma_m))^2\right] ; \quad \text{SE}(\gamma_m) = \sqrt{\text{Var}(\gamma_m)}$$

In ML we typically measure:

$$(\mu_m - 1.96\,SE(\mu_m),\ \mu_m + 1.96\,SE(\mu_m))$$

It corresponds to the 95% of the interval

* MSE as bias/variance

$$\mathbb{E}[(\hat{Y}_m - Y)^2] = \mathbb{E}(\hat{Y}_m^2) - 2\mathbb{E}(\hat{Y}_m)Y + Y^2$$

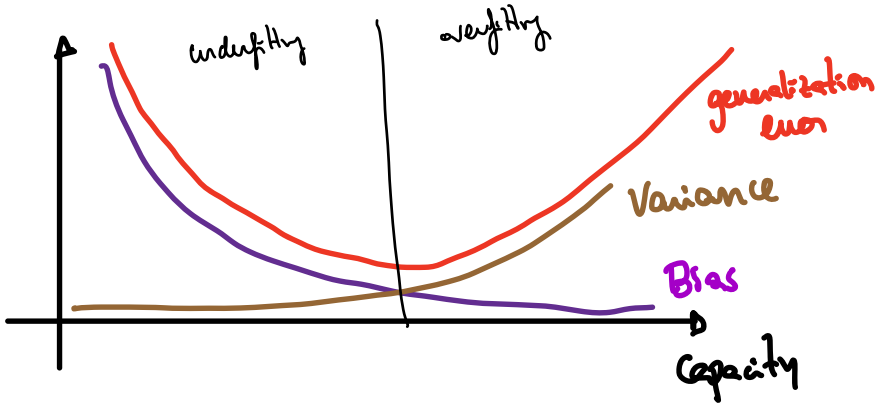
$$Bias^2(\hat{Y}_m, Y) = [\mathbb{E}(\hat{Y}_m) - Y]^2 = \mathbb{E}(\hat{Y}_m)^2 - 2\mathbb{E}(\hat{Y}_m)Y + Y^2$$

$$Var(\hat{Y}_m) = \mathbb{E}(\hat{Y}_m^2) - \mathbb{E}(\hat{Y}_m)^2$$

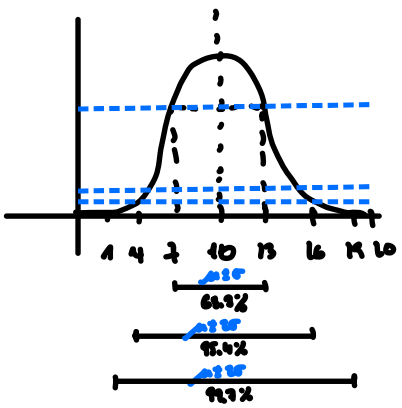
$$Bias^2(\hat{Y}_m, Y) + Var(\hat{Y}_m) = \cancel{\mathbb{E}(\hat{Y}_m)^2} - 2\mathbb{E}(\hat{Y}_m)Y + Y^2 + \mathbb{E}(\hat{Y}_m^2) - \cancel{\mathbb{E}(\hat{Y}_m)^2}$$

$$= \mathbb{E}(\hat{Y}_m^2) - 2\mathbb{E}(\hat{Y}_m)Y + Y^2 = \mathbb{E}[(\hat{Y}_m - Y)^2]$$

$$\boxed{\mathbb{E}[(\hat{Y}_m - Y)^2] = Bias^2(\hat{Y}_m, Y) + Var(\hat{Y}_m)}$$



* **Example** $p(x) = \mathcal{N}(x|\mu,\sigma^2)$ with $\mu = 10$; $\sigma^2 = 9$



Likelihood:

$$p(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

Statistically independent: (iid)

$$p(\vec{x};\mu,\sigma^2) = p(x_1,x_2,\dots,x_m|\mu,\sigma^2)$$
$$= p(x_1;\mu,\sigma^2)p(x_2;\mu,\sigma^2)\dots p(x_m;\mu,\sigma^2)$$
$$= \prod_{i=1}^{m} p(x_i,\mu,\sigma^2) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} = \frac{1}{(2\pi\sigma^2)^{\frac{m}{2}}} e^{-\sum_{i=1}^{m}\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

- **Estimators**
  - Empirical mean: $\hat{\mu}_m = \frac{1}{m}\sum x_i$
  - Empirical variance: $\hat{\sigma}_m^2 = \frac{1}{N}\sum(x_i - \hat{\mu}_m)^2$
  - Unbiased empirical variance: $\bar{\sigma}_m^2 = \frac{1}{N-1}\sum(x_i - \hat{\mu}_m)^2$

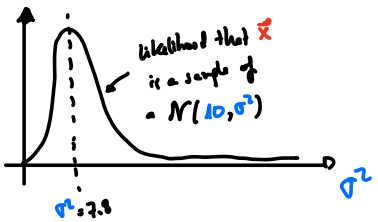
Data:
$x_i = \{10, 12, 7, 5, 11\} \Rightarrow \begin{cases} \hat{\mu}_m = 9 \\ \hat{\sigma}_m^2 = 6.8 \\ \bar{\sigma}_m^2 = 8.5 \end{cases}$

- **We would like:**

$$\begin{cases} E(\hat{\mu}_m) = \mu \Rightarrow \text{small } \text{bias}(\hat{\mu}_m) = E(\hat{\mu}_m) - \mu; \text{ small } \text{Var}(\hat{\mu}_m) = \mathbb{E}[(\mu - \hat{\mu}_m)] \\ E(\hat{\sigma}_m^2) = \sigma^2 \Rightarrow \text{small } \text{bias}(\hat{\sigma}_m^2) = E(\hat{\sigma}_m^2) - \sigma^2: \text{ small } \text{Var}(\hat{\sigma}_m^2) = \mathbb{E}[(\sigma^2 - \hat{\sigma}_m^2)] \\ E(\bar{\sigma}_m^2) = \sigma^2 \Rightarrow \text{small } \text{bias}(\bar{\sigma}_m^2) = E(\bar{\sigma}_m^2) - \sigma^2; \text{ small } \text{Var}(\bar{\sigma}_m^2) = \mathbb{E}[(\sigma^2 - \bar{\sigma}_m^2)] \end{cases}$$

Maximum likelihood estimation (MLE)

* Motivation



$\Rightarrow$ Suppose fixed $\vec{x}$ and $\mu=10$; $\varphi(\vec{x};\mu,\sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} e^{-\sum \frac{1}{2\sigma^2}(x_i-\mu)^2}$

In fact: $\hat{\sigma}^2_m = \frac{1}{N}\sum(x_i-\mu)^2 = 7.8$

Recall:

$$\begin{cases} \mu = \mathbb{E}(X) \\ \sigma^2 = Var(X) = \mathbb{E}[(X-\mu)^2] = \mathbb{E}(X^2) - 2\mu\underbrace{\mathbb{E}(X)}_{\mu} + \mu^2 = \mathbb{E}(X^2)-\mu^2 \\ Var(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \end{cases}$$

- Suppose μ known and call $s = \sigma^2$

$$\begin{cases} \max\limits_{s} \varphi(\vec{x};\mu,s) \Leftrightarrow \frac{\partial \varphi}{\partial s} = 0 \\ \max\limits_{s} \log[\varphi(\vec{x};\mu,s)] \Rightarrow \max \log[\varphi(\vec{x};\mu,s)] = \log\left[\frac{1}{(\sqrt{2\pi s})^n}\right] - \frac{1}{2s}\sum(x_i-\mu)^2 \end{cases}$$

$$-\left(\frac{n}{2}\right)\ln s^{\frac{n}{2}}(2\pi)^{-\frac{n}{2}-1}2\pi + \frac{1}{2s^2}\sum(x_i-\mu)^2 = 0 \Rightarrow -\left(\frac{n}{2}\right)\frac{2\pi}{(2\pi)s} + \frac{1}{2s^2}\sum(x_i-\mu)^2 = 0 \Leftrightarrow s = \boxed{\hat{\sigma}^2_m = \frac{1}{N}\sum(x_i-\mu)^2}$$

$$\mathbb{E}(s) = \mathbb{E}(\hat{\sigma}^2_m) = \frac{1}{m}\sum \mathbb{E}[(x_i-\mu)^2] = \frac{1}{m}\sum\left[\mathbb{E}(x_i^2) - 2\mu\underbrace{\mathbb{E}(x_i)}_{\mu} + \mu^2\right] = \frac{1}{m}\sum\left[\mathbb{E}(x_i^2)-\mu^2\right] = \sigma^2+\mu^2-\mu^2 = \sigma^2$$

$$\boxed{\mathbb{E}(\hat{\sigma}^2_m) = \sigma^2}$$

- Suppose μ, σ^2 unknown

$$\arg\max_{\mu,s} \varphi(\vec{x};\mu,s) \Leftrightarrow \arg\max_{\mu,s} \log[\varphi(\vec{x};\mu,s)] \Rightarrow \left\{ \begin{array}{l} \frac{\partial \log \varphi(\vec{x};\mu,s)}{\partial \mu} = 0 \\ \frac{\partial \log \varphi(\vec{x};\mu,s)}{\partial s} = 0 \end{array} \right\} \Rightarrow (\mu_m, s_m)$$

$$* \quad \frac{\partial \log \varphi(\vec{x};\mu,s)}{\partial \mu} = \frac{\partial}{\partial \mu}\log\left[\frac{1}{(\sqrt{2\pi s})^n} e^{-\sum \frac{1}{2s}(x_i-\mu)^2}\right] = \frac{\partial}{\partial \mu}\left[-\log[(\sqrt{2\pi s})^n] - \sum \frac{1}{2s}(x_i-\mu)^2\right] = -\frac{1}{s}\sum(x_i-\mu)\cdot(-1) = \frac{1}{s}\left[\sum x_i - m\mu\right] = 0 \Rightarrow \mu = \frac{1}{m}\sum x_i$$

$$\mathbb{E}(\mu_m) = \frac{1}{m}\sum \mathbb{E}(x_i) = \frac{1}{m} m\mu = \mu \Rightarrow \boxed{\mathbb{E}(\mu_m) = \mu}$$

$$* \quad \frac{\partial \log \varphi(\vec{x};\mu,s)}{\partial s} = \frac{\partial}{\partial s}\left[-\log[(\sqrt{2\pi s})^n] - \sum \frac{1}{2s}(x_i-\mu)^2\right] = -\frac{1}{(\sqrt{2\pi s})^n}\, n/2\,(2\pi)^{\frac{n}{2}-1} - \sum -\frac{1}{2s^2}(x_i-\mu)^2 = -\frac{n/2\,m}{(2\pi s)} + \frac{1}{2s^2}\left[\sum (x_i-\mu)^2\right]$$

$$= -\frac{m}{2s} + \frac{1}{2s^2}\left[\sum (x_i-m\mu)^2\right] = \frac{m}{2s}\left[\frac{1}{ms}\left[\sum (x_i-m\mu)^2\right]^1 - 1\right] = 0 \Rightarrow s_m = \frac{1}{m}\left[\sum (x_i-\mu)^2\right]$$

$$s_m = \frac{1}{m}\sum (x_i-\mu)^2 = \frac{1}{m}\left[\sum x_i^2 - 2\sum x_i\mu + \sum \mu^2\right] = \frac{1}{m}\left[\sum x_i^2 - 2\sum x_i \frac{1}{m}\sum x_i + \sum \left(\frac{1}{m}\sum x_i\right)^2\right]$$

$$= \frac{1}{m}\left[\sum x_i^2 - \frac{2}{m}\left(\sum x_i\right)^2 + \frac{1}{m^2}\underbrace{\sum \left(\sum x_i\right)^2}_{m\left(\sum x_i\right)^2}\right] = \frac{1}{m}\left[\sum x_i^2 - \frac{1}{m}\left(\sum x_i\right)^2\right]; \quad \boxed{\sigma^2_m = \frac{1}{m}\left[\sum x_i^2 - \frac{1}{m}\left(\sum x_i\right)^2\right]}$$

$$\mathbb{E}\left[\left(\sum x_i\right)^2\right] = \mathbb{E}\left[\sum x_i^2 + \sum_i \sum_{j\neq i} x_i x_j\right] = m\,\mathbb{E}(x_i^2) + \sum_i \sum_{j\neq i} \mathbb{E}(x_i)\mathbb{E}(x_j) = m(\sigma^2+\mu^2) + \sum_i \mathbb{E}(x_i)(m-1)\mu = m(\sigma^2+\mu^2) + m(m-1)\mu^2 = m[\sigma^2 + m\mu^2]$$

$$\mathbb{E}(\sigma^2_m) = \frac{1}{m}\mathbb{E}\left[\sum x_i^2 - \frac{1}{m}\left(\sum x_i\right)^2\right] = \frac{1}{m}\left[\underbrace{\mathbb{E}\left(\sum x_i^2\right)}_{m(\sigma^2+\mu^2)} - \frac{1}{m}\underbrace{\mathbb{E}\left[\left(\sum x_i\right)^2\right]}_{\frac{1}{m}[\sigma^2+m\mu^2]}\right] = \frac{1}{m}\left[m(\sigma^2+\mu^2) - \frac{1}{m}m[\sigma^2+m\mu^2]\right] = \frac{1}{m}\left[(m-1)\sigma^2 + (m-m)\mu^2\right] = \frac{m-1}{m}\sigma^2$$

$$\boxed{\mathbb{E}(\sigma^2_m) = \frac{m-1}{m}\sigma^2}$$

MLE: Maximum likelihood estimation

- **Idea**: Pick θ that assign the maximum probability to the training data D
- **Assumption**: iid assumption: $p(D|\theta) = \prod_{n=1}^{N} p(y_n|x_n;\theta)$
- **Definition**: $\hat{\theta}_{mle} = \arg\max_\theta p(D|\theta) = \arg\max_\theta \log(p(D|\theta))$

  Thus, $\hat{\theta}_{mle} = \arg\max_\theta \sum_{i=1}^{n} \log(p(y_n|x_n;\theta)) = \arg\min_\theta \underbrace{-\sum_{i=1}^{n} \log(p(y_n|x_n;\theta))}$

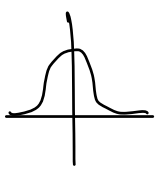
  (conditional) Negative log likelihood (NLL)

Example: MLE for linear regression

$p(y|\theta) = \mathcal{N}(y|\theta^T x, \sigma^2)$. Assume σ^2 fixed

$$\arg\min NLL(\theta) = -\sum_{n=1}^{N} \log\left[\frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2}(y_n-\theta^T x)^2}\right]$$

$$= \sum_{n=1}^{N} \log(2\pi\sigma^2)^{1/2} + \frac{1}{2\sigma^2}(y_n-\theta^T x)^2 \propto$$



$$\arg\min RSS(\theta) = \sum_{n=1}^{N} (y_n-\theta^T x)^2 \quad \text{residual sum of squares} =$$

$$\arg\min MSE(\theta) = \frac{1}{N} RSS(\theta) = \frac{1}{N}\sum_{n=1}^{N}(y_n-\theta^T x)^2$$ Ordinary least square

$$\Rightarrow \frac{1}{N}(X\omega - y)^T(X\omega - y) \Rightarrow \theta^* = (X^T X)^{-1} X^T y$$

Example Logistic regression model

Given $y \in \{0,1\}$, Bernulli: $Pr(y|X;\theta) = g(h_\theta(x))^y (1-g(h_\theta(x)))^{1-y}$

Likelihood (iid): $L(\theta|y;x) = Pr(y|X;\theta) = \prod_i Pr(y_i|x_i;\theta)$

$$= \prod_i^m g(h_\theta(x_i))^{y_i} (1-g(h_\theta(x_i)))^{1-y_i}$$

$$\arg\max_\theta \log\left[\prod_i^m g(h_\theta(x_i))^{y_i} (1-g(h_\theta(x_i)))^{1-y_i}\right] =$$

$$\arg\min_\theta \frac{1}{m}\sum_{i=1}^{m} (1-y^i)(-\log(1-g(h_\theta(x^i)))) + y^i(-\log(g(h_\theta(x^i))))$$

Maximum a posteriori probability estimation (MAP)

* Bayesian version of MLE + a prior probability density on the parameter θ

* θ : vector to be estimated
  $D = \{x, y\}$: observations
  } are random variables with joint probability density $p(D, \theta)$

* Statistical estimation vs Bayesian

  θ is an unknown parameter
  $\hat{\theta}$ estimate is random since D is seen as random

  θ is a random variable
  D is observed (not random)

* Prior density: $p(\theta) = \int p(D, \theta)\, dD$ } Prior info about vector θ before we observe D

  $p(D) = \int p(D, \theta)\, d\theta$ } Prior info about the measurements or observations D.

* Conditional density of D given θ: $p(D|\theta) = \frac{p(D, \theta)}{p(\theta)}$

  Plays the role of probability in the MLE

* Conditional density of θ given D: $\underbrace{p(\theta|D)}_{\text{Posterior density of } \theta} = \frac{p(D, \theta)}{p(D)} = p(D|\theta)\frac{p(\theta)}{p(D)}$

  = Bayes rule"

Bayes theorem

joint probability : $p(H, E)$ — E = evidence, H = hypothesis

prior: $P(H)$ = probability an hypothesis is true (before any evidence); $P(E)$ = probability of seeing the evidence

likelihood: $P(E|H)$ = probability of seeing the evidence if hypothesis is true

Posterior: $p(H|E)$ = probability an hypothesis is true given some evidence

$p(D|\theta)$

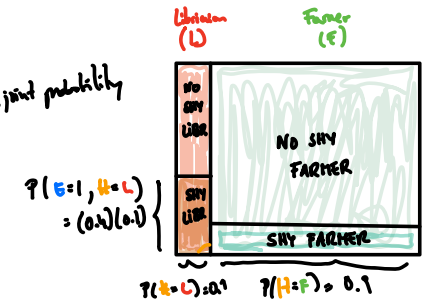
Bayes theorem:

$$p(H|E) = P(E|H)\frac{P(H)}{P(E)}$$

* Farmers/librarian example

	$P(H=L) = 0.1$	$P(H=F) = 0.9$
$P(E=0)$ 0.87	$P(E=0, H=L)$ $(0.1)(0.1)$	$P(E=0, H=F)$ $(0.9)(0.9)$
$P(E=1)$ 0.13	$P(E=1, H=L)$ $(0.4)(0.1)$	$P(E=1, H=F)$ $(0.1)(0.9)$

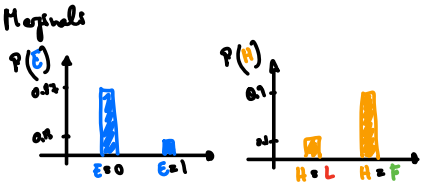
joint probability



$P(E=1, H=L) = (0.4)(0.1)$

$P(E=1, F) = (0.1)(0.1)$

Marginals



\+ Likelihood (Conditional probability)

a) Probability of being shy $E=1$ when being a librarian $H=L$

$$p(E=1|H=L) = \frac{P(E=1, H=L)}{P(H=L)} = 0.4$$

b) Probability of being shy $E=1$ when being a farmer $H=F$

$$p(E=1|H=F) = \frac{P(E=1, H=F)}{P(H=F)} = 0.1$$

* Posterior

a) Probability of being a librarian $H=L$ when being shy $E=1$

$$p(H=L|E=1) = \frac{P(E=1, H=L)}{P(E=1)} = \frac{(0.4)(0.1)}{0.13} = 0.307$$

b) Probability of being a farmer $H=F$ when being shy $E=1$

$$p(H=F|E=1) = \frac{P(E=1, H=F)}{P(E=1)} = \frac{(0.1)(0.9)}{0.13} = 0.692$$

Remark

Typically the interest is to know the posterior $p(H|E=1)$ by knowing $p(E=1|H)$ and $P(H)$ and $P(E=1)$

- considering that we know the probability of being shy when is a librarian $p(E=1|H=L)$, knowing the probability of being a librarian $P(H=L)$ and the probability of being shy $P(E=1)$, the

¿What is the probability to be a librarian if the person is shy? i.e $p(H=L|E=1)$?

* Our estimate of θ, given the observation D:

We take as estimate of θ, the value that maximizes the conditional density of θ given the observed value of D

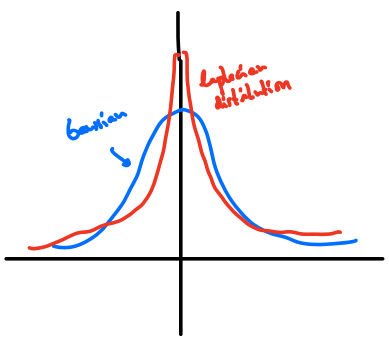
$$\hat{\theta}_{map} = \arg\max_{\theta} \underbrace{p(\theta|D;\theta,D)}_{\text{post}} = \arg\max_{\theta} \underbrace{p(D|\theta;\theta,D)}_{\text{likelihood}} \underbrace{p_X(\theta)}_{\text{prior}} = \arg\max_{\theta} p(\theta,D)$$

* The difference with MLE is $p_X(\theta)$ is now considered (we consider the prior info of x)

{If the prior is uniform MAP ≡ MLE)

* Taking logarithms: $\hat{\theta}_{map} = \arg\max_{\theta} \underbrace{\log(p(D|\theta;\theta,D))}_{\text{Same term as the log-likelihood}} + \underbrace{\log(p_X(\theta))}_{\text{Penalization of choices of } \theta \text{ that are very unlikely to happen}}$

$$= \arg\min_{\theta} -\log(p(D|\theta;\theta,D)) - \log(p_X(\theta))$$



Examples:

Gaussian with zero mean: $p_X(\theta) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(\frac{-\theta^2}{2\sigma^2}\right)$

$-\log[p_X(\theta)] \simeq \frac{\theta^2}{2\sigma^2}$ l^2 regularization

Laplacian: $p_X(\theta) = \frac{1}{(2a)} \exp\left(\frac{|\theta|}{a}\right)$

$-\log[p_X(\theta)] \simeq \frac{|\theta|}{a}$ l^1 regularization