- Previously assumed we knew priors $P_Y(y)$ and $f_{\underline{X}|Y}(\underline{x}|y)$
- But rarely have complete knowledge
- May have general knowledge plus samples / training data
- Use to design / train classifier.
- Use samples to estimate unknown probabilities / pdfs and use estimates as true values.
- Estimation of $f_{\underline{X}|Y}(\underline{x}|y)$ problematic when # samples too small and when dimensionality of $\underline{x}$ is large
- If can parameterize conditional PDF, e.g. assume $f_{\underline{X}|Y}(\underline{x}|i) \sim \mathcal{N}(\underline{\mu}_i, \Sigma_i)$ (without knowing $\underline{\mu}_i, \Sigma_i$) simplifies problem to parameter estimation
- Maximum Likelihood Estimation — parameters with fixed unknown values: best estimate is one which maximizes prob. of obtaining observed samples.
- Bayesian methods: parameters as r.v.'s having known prior distribution.

  Observation of samples converts this to a posterior density. Typically, additional samples sharpen a posterior density, peak near true values

ML estimation

- have good convergence properties as # training samples increases
- simpler than alternative methods
- use set D of training samples drawn indep. from $f(\underline{x}|\underline{\theta})$ to estimate unknown parameter $\underline{\theta}$
  e.g. $\underline{\theta} = (\underline{\mu}, \Sigma)$
- suppose D contains samples $\underline{x}_1, \ldots, \underline{x}_n$

$$f(D|\underline{\theta}) = \prod_{k=1}^{n} f(\underline{x}_k | \underline{\theta})$$

<u>likelihood</u> of $\underline{\theta}$ wrt ~~same~~ set of samples

- ML estimate of $\underline{\theta}$ is $\underline{\hat{\theta}}_{ML}$

$$\underline{\hat{\theta}}_{ML} = \underset{\underline{\theta}}{\operatorname{argmax}}\ f(D|\underline{\theta})$$

value of $\underline{\theta}$ that best agrees/supports actually observed training samples.

- $\ln(\cdot)$ ↗ usu. easier to work with

<u>log likelihood</u> function

$$\underline{\hat{\theta}}_{ML} = \underset{\underline{\theta}}{\operatorname{argmax}}\ \ln f(D|\underline{\theta})$$

$$= \underset{\underline{\theta}}{\operatorname{argmax}}\ \ln \prod_{k} f(\underline{x}_k|\underline{\theta})$$

$$= \underset{\underline{\theta}}{\operatorname{argmax}} \sum_{k=1}^{n} \ln f(\underline{x}_k|\underline{\theta})$$

Ex: <u>Data</u>: observed seq. D of n_H heads and n_T tails

<u>Model</u>: Each flip follows Bernoulli distribution

$$P(H) = \theta, \quad P(T) = 1-\theta, \quad \theta \in [0,1].$$

likelihood of observing seq. D is

$$P(D|\theta) = \theta^{n_H} (1-\theta)^{n_T}$$

Given model and data, estimate θ.

<u>ML estimate</u>:
$$\begin{aligned}
\hat{\theta}_{ML} &= \underset{\theta}{\operatorname{argmax}}\ P(D|\theta) \\
&= \underset{\theta}{\operatorname{argmax}}\ \ln P(D|\theta) \\
&= \underset{\theta}{\operatorname{argmax}}\ \ln\left(\theta^{n_H}(1-\theta)^{n_T}\right) \\
&= \underset{\theta}{\operatorname{argmax}}\ \underbrace{n_H \ln\theta + n_T \ln(1-\theta)}_{\text{concave fn. of } \theta}
\end{aligned}$$

$$\frac{d}{d\theta} \ln P(D|\theta) = 0:$$

$$\frac{n_H}{\theta} - \frac{n_T}{1-\theta} = 0$$

$$\Rightarrow \hat{\theta}_{ML} = \frac{n_H}{n_H + n_T}$$

So e.g. flip 10 times, get H 8 times, T 2 times

Then $\hat{\theta}_{ML} = \dfrac{n_H}{n_H + n_T} = 0.8$

Trusting data completely. There could be too little or noisy data.

Ex: Data: observe $D = \{x_1, \ldots, x_n\}$

Model: $X_i \sim \mathcal{N}(\mu, \sigma^2)$ iid

Estimate: μ, σ^2

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \ln f(D \mid \theta)$$

$$= \quad '' \quad \ln \prod_{k=1}^{n} f(x_k \mid \theta)$$

$$= \quad '' \quad \ln \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_k-\mu)^2}{2\sigma^2}\right]$$

$$= \quad '' \quad \sum_{k=1}^{n} -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \frac{(x_k-\mu)^2}{2\sigma^2}$$

$$= \underset{\theta}{\operatorname{argmin}} \sum_{k=1}^{n} \frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln\sigma^2 + \frac{(x_k-\mu)^2}{2\sigma^2}$$

convex in μ

$$\frac{\partial \ln f(D \mid \theta)}{\partial \mu} = \sum_{k=1}^{n} \frac{(x_k-\mu)}{\sigma^2} = 0$$

$$\Rightarrow \hat{\mu}_{ML} = \frac{1}{n}\sum_{k=1}^{n} x_k \quad \text{sample mean}$$

$$\frac{\partial \ln f(D \mid \theta)}{\partial \sigma^2} = 0: \quad \hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{k=1}^{n} (x_k - \hat{\mu}_{ML})^2$$

Is $\hat{\mu}_{ML}$ unbiased?

$E[\hat{\mu}_{ML}] \stackrel{?}{=} \mu$ i.e. expected value of estimator over all data sets of size n = true value

$$E[\hat{\mu}_{ML}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_k\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_k] = \frac{1}{n}\sum_{i=1}^{n} \mu$$

$$= \mu \checkmark$$

unbiased

Is $\hat{\sigma}^2_{ML}$ unbiased?

$$E[\hat{\sigma}^2_{ML}] = E\left[\frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu}_{ML})^2\right]$$

$$= E\left[\frac{1}{n}\sum_{k=1}^{n}((x_k-\mu)+(\mu-\hat{\mu}))^2\right]$$

$$= E\left[\frac{1}{n}\sum_{k=1}^{n}\{(x_k-\mu)^2 - 2(x_k-\mu)(\hat{\mu}-\mu)+(\hat{\mu}-\mu)^2\}\right]$$

$$= E\left[\frac{1}{n}\sum_{k=1}^{n}(x_k-\mu)^2\right] - 2E\left[\frac{1}{n}\sum_{k=1}^{n}(x_k-\mu)(\hat{\mu}-\mu)\right] + E\left[\frac{1}{n}\sum_{k=1}^{n}(\hat{\mu}-\mu)^2\right]$$

$$= \frac{1}{n}\sum_{k=1}^{n}\sigma^2 \qquad -2E\left[\left(\underbrace{\frac{1}{n}\sum_{k=1}^{n}x_k}_{} - \frac{1}{n}\sum_{k=1}^{n}\mu\right)(\hat{\mu}-\mu)\right] \qquad + E[(\hat{\mu}-\mu)^2]$$

$$= \sigma^2 \qquad -2E[(\hat{\mu}-\mu)(\hat{\mu}-\mu)]$$

$$\underbrace{-2E[(\hat{\mu}-\mu)^2]}$$

$$-\underbrace{E[(\hat{\mu}-\mu)^2]}$$

$$\mathrm{var}\left(\frac{1}{n}\sum_{k=1}^{n}x_i\right)$$

$$= \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

$$E[\hat{\sigma}^2_{ML}] = \sigma^2 - \frac{\sigma^2}{n}$$

$$= \frac{(n-1)\sigma^2}{n}$$

<u>biased</u>

However $\lim_{n\to\infty} E[\hat{\sigma}^2_{ML}] = \sigma^2$

<u>asymptotically</u> <u>unbiased</u> : estimator becomes unbiased in limit as # samples $\to \infty$

Properties of ML estimators

- Assume data generated from some true parameter value $X \sim f(\underline{x} \mid \theta^*)$. ML estimator $\hat{\theta}_{ML}$ obtained from data set of size n satisfies.

  - asymptotically unbiased $\lim_{n \to \infty} E[\hat{\theta}_{ML}] = \theta^*$ ~~as n→∞~~
  - variance $(\hat{\theta}_{ML})$ is minimum among estimators
  - If $\hat{\theta}_{ML}$ is for θ, then $g(\hat{\theta}_{ML})$ is ML estimator for $g(\theta)$.
  - Distribution of $\hat{\theta}_{ML}$ is Gaussian for large n.

Multivariate Gaussian

$$D = (\underline{x}_1, \ldots, \underline{x}_n)$$

$$f(D \mid \mu, \Sigma) = \prod_{k=1}^{n} f(\underline{x}_k \mid \underline{\mu}, \Sigma)$$

$$= \prod_{k=1}^{n} \frac{1}{(2\pi)^{d/2} \sqrt{\det(\Sigma)}} \exp\left[-\frac{1}{2} (\underline{x}_k - \underline{\mu})^T \Sigma^{-1} (\underline{x}_k - \underline{\mu}) \right]$$

$$\ln f(D \mid \mu, \Sigma) = \sum_{k=1}^{n} -\frac{1}{2} \ln\left[(2\pi)^d \det(\Sigma) \right] - \frac{1}{2} (\underline{x}_k - \underline{\mu})^T \Sigma^{-1} (x_k - \underline{\mu})$$

$$\frac{\partial}{\partial \underline{\mu}} = \underline{0} \qquad \hat{\underline{\mu}}_{ML} = \frac{1}{n} \sum_{k=1}^{n} \underline{x}_k$$

$$\frac{\partial}{\partial \Sigma} = 0: \qquad \hat{\Sigma}_{ML} = \frac{1}{n} \sum_{k=1}^{n} (\underline{x}_k - \hat{\underline{\mu}}_{ML})(\underline{x}_k - \hat{\mu}_{ML})^T$$

MAP Estimation

- $\underline{\theta}$ considered as a random vector; Have prior belief regarding $p(\underline{\theta})$

$$\hat{\underline{\theta}}_{MAP} = \underset{\underline{\theta}}{\operatorname{argmax}}\ p(\underline{\theta} \mid D)$$

incorporate data $D = \{\underline{x}_1, \dots, \underline{x}_n\}$

$$= \underset{\underline{\theta}}{\operatorname{argmax}}\ \frac{p(D \mid \underline{\theta})\, p(\underline{\theta})}{p(D)}$$

$$= \underset{\underline{\theta}}{\operatorname{argmax}}\ \underbrace{p(D \mid \underline{\theta})}_{\text{likelihood}}\ \underbrace{p(\underline{\theta})}_{\text{prior}}$$

$$= \underset{\underline{\theta}}{\operatorname{argmax}} \left(\prod_{k=1}^{n} p(\underline{x}_k \mid \underline{\theta}) \right) p(\underline{\theta})$$

$$= \underset{\underline{\theta}}{\operatorname{argmax}} \sum_{k=1}^{n} \left(\ln p(\underline{x}_k \mid \underline{\theta}) + \ln p(\underline{\theta}) \right)$$

Ex: Assume $f(\underline{x} \mid \mu) = \mathcal{N}(\mu, \underset{\text{known}}{\sigma^2})$ where μ is a r.v. with prior $f(\mu)$

$$\underbrace{f(\mu)}_{\text{prior}} = \mathcal{N}(\mu_0, \sigma_0^2) \quad \text{known}$$

μ_0 : best prior guess for μ

σ_0^2 : uncertainty in prior guess for μ

$D = (x_1, \dots, x_n)$ are observed samples

$$\hat{\mu}_{MAP} = \underset{\mu}{\operatorname{argmax}}\ \underbrace{f(\mu \mid D)}_{\text{posterior}}$$

$$= \underset{\mu}{\operatorname{argmax}} \; f(D \mid \mu) \, f(\mu)$$

$$= \;'' \; \prod_{k=1}^{n} f(x_k \mid \mu) \, f(\mu)$$

$$= \;'' \; \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}}$$

$$= \;'' \; -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2$$

$$-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma_0^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2$$

$$\frac{\partial}{\partial\mu} = 0: \qquad \frac{1}{\sigma^2}\sum_{k=1}^{n}(x_k-\mu) = \frac{1}{\sigma_0^2}(\mu-\mu_0)$$

$$\hat{\mu}_{MAP} = \frac{\sigma_0^2}{\sigma^2+\sigma_0^2 n}\sum_{k=1}^{n} x_k + \frac{\sigma^2}{\sigma^2+\sigma_0^2 n}\mu_0$$

$$= \frac{\sigma_0^2 n}{\sigma^2+\sigma_0^2 n}\underbrace{\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)}_{\substack{\hat{\mu}_{ML} \\ \text{data}}} + \frac{\sigma^2}{\sigma^2+\sigma_0^2 n}\underbrace{\mu_0}_{\text{prior}}$$

- Note $\hat{\mu}_{MAP}$ bet. μ_0 and $\hat{\mu}_{ML}$

$$\underset{\text{prior}}{\underset{\uparrow}{\mu_0}} \qquad \underset{\text{data}}{\underset{\uparrow}{\hat{\mu}_{ML}}}$$

- If $\sigma_0 \neq 0$, as $n \to \infty$, $\hat{\mu}_{MAP} \to \hat{\mu}_{ML}$
- If $\sigma_0 = 0$, $\hat{\mu}_{MAP} = \mu_0$
  - degenerate situation where prior is certain

- If $\sigma_0^2 \gg \sigma$, $\hat{\mu}_{MAP} \sim \hat{\mu}_{ML}$: high uncertainty about prior
- If σ^2/σ_0^2 is finite, after enough samples $\hat{\mu}_{MAP}$ will converge to $\hat{\mu}_{ML}$ regardless of μ_0 and σ_0^2.

<u>Ex</u>: coin flip example: $D = \{ \alpha_1 \; H, \; \alpha_0 \; T\text{'s} \}$

suppose have prior $f(\theta) = \text{Beta}(\beta_0, \beta_1) = \dfrac{\theta^{\beta_1 - 1}(1-\theta)^{\beta_0 - 1}}{\underbrace{B(\beta_0,\beta_1)}_{\text{normalization}}}$

β_1, β_0 indicate fractions and H and T.

$$\hat{\theta}_{MAP} = \arg\max_\theta \; P(D \mid \theta) f(\theta)$$

$$= \;'' \quad \theta^{\alpha_1}(1-\theta)^{\alpha_0} \frac{\theta^{\beta_1 - 1}(1-\theta)^{\beta_0 - 1}}{B(\beta_0,\beta_1)}$$

$$= \;'' \quad \frac{\theta^{\alpha_1 + \beta_1 - 1}(1-\theta)^{\alpha_0 + \beta_0 - 1}}{B(\beta_0,\beta_1)}$$

$$= \;'' \quad \theta^{\alpha_1 + \beta_1 - 1}(1-\theta)^{\alpha_0 + \beta_0 - 1}$$

$$= \frac{(\alpha_1 + \beta_1 - 1)}{(\alpha_1 + \beta_1 - 1) + (\alpha_0 + \beta_0 - 1)}$$

Add $\beta_1 - 1$ imag. heads
$\beta_0 - 1$ imag tails

$$\text{vs.} \quad \hat{\theta}_{ML} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

why choose beta prior, b/c a posterior is also beta

<u>conjugate prior</u>