PRML 1. INTRODUCTION

- Supervised learning { classification ... discrete output
  regression ... continuous output
- Unsupervised learning { clustering
  density estimation
  visualization } etc.
- Reinforcement learning : exploration ⟷ exploitation

§ 1.1

"regularization"
("shrinkage" or "weight decay")

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_n \left[y(x_n, \mathbf{w}) - t_n\right]^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

expectation, data, $\sim 10^{-6}$

quadratic ⇒ "ridge regression"

const. term can be omitted.

§ 1.2

Bayes' theorem $p(Y|X) = \dfrac{p(X|Y)p(Y)}{p(X)} = \dfrac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)}$

$\left(= \dfrac{p(X,Y)}{p(X)}\right)$

Maximum likelihood : a frequentists' estimator $p(D|\mathbf{w})$ (or $-\ln p(D|\mathbf{w})$)

for given params

the prob. to obtain data D

[cf. "bootstrap" method for uncertainty of $\mathbf{w}$]

Gaussian distribution

$$N(\mathbf{x}\mid\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

maximal likelihood $\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$

$$\sigma^2_{ML} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})^2$$

but, $\mathbb{E}(\mu_{ML}) = \mu$, $\mathbb{E}(\sigma^2_{ML}) = \frac{N-1}{N}\sigma^2$: biased

- Curve-fitting: predict $t = y(x,w) = \sum_{i=0}^{M} w_i x^i$

from noised-data $\{x, t\}$ assuming Gaussian noise β

[i.e. $p(t\mid x, w, \beta) = N(t\mid y(x,w), \beta^{-1})$] ··· (2)

⇒ likelihood function

$$\ln p(t\mid x, w, \beta) = -\frac{\beta}{2}\sum\left[y(x_i,w)-t_i\right]^2 + \frac{N}{2}\ln\beta + \text{const.} \quad \cdots (1)$$

→ w_{ML} and β_{ML} are obtained $\frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}_{ML})-t_n\}^2$.

We can introduce a prior. e.g. $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0},\alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2}\exp\left\{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right\}$

(α: hyperparameter: parameter controlling the models.)

⇒ likelihood

$$p(w\mid x, t, \beta, \alpha) = \frac{p(t\mid x, w, \beta)\, p(w\mid\alpha)}{\sum_{w}(\text{numerator})} \quad \cdots (3)$$

($p(w\mid x,t,\beta,\alpha)$: posterior on w; $p(t\mid x,w,\beta)$: new information; $p(w\mid\alpha)$: prior on w)

$$\ln p(w\mid x, t, \beta, \alpha) = (1) - \frac{\alpha}{2}\|w\|^2 + \text{const.}$$

··· w_{ML} is given by "regularized" $\tilde{E}$ in §1.1.

with $\lambda = \alpha/\beta$.

(but this is "point estimation" ... not Bayesian way)

• Curve-fitting in Bayesian way

new point, data

Bayesian approach: consider $p(t \mid x, \mathbf{x}, \mathbf{t})$

prediction

$$p(t \mid x, \mathbf{x}, \mathbf{t}) = \int p(t \mid x, w)\, p(w \mid \mathbf{x}, \mathbf{t})\, dw$$

we knew (2) (3) [α and β are assumed fixed]

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}\left(t|m(x), s^2(x)\right) \tag{1.69}$$

where the mean and variance are given by

dependent on $\mathbf{x}$ (!)

$\Rightarrow$

$$m(x) = \beta \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n \tag{1.70}$$

$$s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \boldsymbol{\phi}(x). \tag{1.71}$$

we had in previous approach

new contribution to the uncertainty.

Here the matrix $\mathbf{S}$ is given by

$$\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} \tag{1.72}$$

where $\mathbf{I}$ is the unit matrix, and we have defined the vector $\boldsymbol{\phi}(x)$ with elements $\phi_i(x) = x^i$ for $i = 0, \ldots, M$.

§ 1.3 Model Selection

- separate data into training + validation (+ test) set
- cross validation: S-times run w. $\frac{S-1}{S}$ training + 1 validation
  (leave-one-out for $S = N$)
  - heavy calculation
  - difficult for multi-parameter model selection
- Akaike information criterion $\ln p(\mathcal{D} \mid w_{\mathrm{ML}}) - M$
- Bayesian information criterion

prefers overly-simple models

§ 1.4 Curse of dimensionality

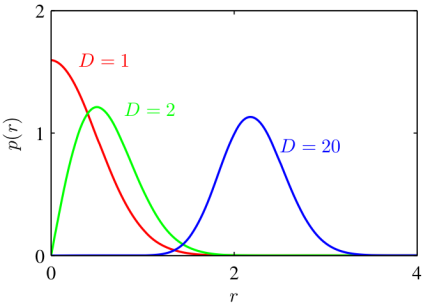


Figure 1.23 Plot of the probability density with respect to radius r of a Gaussian distribution for various values of the dimensionality D. In a high-dimensional space, most of the probability mass of a Gaussian is located within a thin shell at a specific radius.

§ 1.5 Decision theory

$$p(C_k|x) = \frac{p(x|C_k)\,p(C_k)}{p(x)}$$

($p(C_k)$: prior; C_k: "class"; x: observed data)

- to assign x to a class $\Leftrightarrow$ to separate the space of x into regions R_k

$$p(\text{correct}) = \sum_k \int_{R_k} dx\; p(x, C_k)$$

- to maximise $p(\text{correct})$: k for x is what gives the largest $p(x, C_k)$ $\Leftrightarrow$ " $p(C_k|x)$

- to minimize "expected loss" L_{kj} (true class, assigned class)

$$\mathbb{E}(L) = \sum_{k,j} \int_{R_j} L_{kj}\, p(x, C_k)\, dx$$

$\Leftrightarrow$ choose j that minimizes $\sum_k L_{kj}\, p(x, C_k)$

$\left(\Leftrightarrow \sum_k L_{kj}\, p(C_k|x)\right)$

- "reject" option (postpone our decision)

$$\max_k p(C_k|x) < \theta$$

(Note that this is a good criterion. Also θ should be $\in (1/k, 1)$)

- Classification problem = Inference + Decision.

(a) calculate $p(C_k, x)$, $p(x)$, etc. and then decide
- too much computation
- very useful to detect rare cases [outlier detection, novelty detection]

(b) calculate $p(C_k|x)$ and then decide.

(c) find a function $f : x \mapsto k$
- no access to posterior probabilities.

Decision theory for Regression

$$\mathbb{E}[L] = \int dx\, dt\; L(t, y(x))\, p(x,t)$$

$y(x)$: estimate, $\mathbb{R}^N \to \mathbb{R}$; x: input, $\in \mathbb{R}^N$; t: true, $\in \mathbb{R}$

for $L = [t - y(x)]^2$,

$$\frac{\delta \mathbb{E}[L]}{\delta y(x)} = 2 \int dx\, dt\; [y(x) - t]\, p(x,t)$$

$$\therefore y(x) = \frac{\int dt\, t\, p(x,t)}{p(x)} = \int dt\, t\, p(t|x) \equiv \mathbb{E}_t[t|x]$$

$\mathbb{E}_t$: for averaged t; $[t|x]$: under the condition x; expectation of t

[OR: $\mathbb{E}[L] = \int dx\, dt \left\{ [y(x) - \mathbb{E}[t|x]]^2 p(x,t) \right.$

$+ 2\big(y(x) - \mathbb{E}[t|x]\big)\big(\mathbb{E}[t|x] - t\big) p(x,t)$ → 0

$\left. + [\mathbb{E}[t|x] - t]^2 p(x,t) \right\}$

$\mathbb{E}[t|x] p(x) \to \mathbb{E}\, p(x)$

dep. on x, indep. of t

$$= \int dx\, [y(x) - \mathbb{E}]^2 p(x) + \int dx\, dt\, [\mathbb{E}[t|x] - t]^2 p(x,t)$$

$p(x,t) = p(x)\, p(t|x)$

$$\hookrightarrow \int dx\; \mathrm{Var}_t[t|x]\, p(x)$$

- $\mathbb{E}_t[f(t)|x] = \int dt\; f(t)\, p(t|x)$
- $\mathbb{V}[Y|X] = \mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2 \,\middle|\, X \right]$
- $\mathbb{V}_t[f(t)|x] = \mathbb{E}_t\left[(f(t) - \mathbb{E}[f(t)|x])^2 \,\middle|\, x \right]$

$$= \int dt\; (f(t) - \mathbb{E}[f(t)|x])^2\, p(t|x)$$

§ 1.6 Entropy

discrete: $H = -\sum_{i=1}^{N} p_i \ln p_i \quad \in [0, \ln N]$

continuous: $H[x] = -\int p(x) \ln p(x)\, dx$

$$\begin{aligned} H_{\text{disc.}} &= -\sum p_i \ln p_i \\ &\to -\sum p(x)\Delta x \ln p(x)\Delta x \\ &= -\sum p(x) \ln p(x)\, \Delta x - \sum p(x) \ln \Delta x\, \Delta x \\ &= -\int p(x) \ln p(x)\, dx - \lim_{\Delta \to 0} \ln \Delta \\ &= H[x] + \infty \end{aligned}$$

for mean μ and variance σ^2,

Gaussian distrib. gives maximal entropy $\frac{1 + \log 2\pi\sigma^2}{2}$

- Conditional entropy $H[y|x] = -\int p(x,y) \ln p(y|x)\, dx\, dy$

$$= H[x,y] - H[x]$$

- Relative entropy (Kullback-Leiber divergence)

$$KL(p \| q) = -\int p(x) \ln \frac{q(x)}{p(x)}\, dx$$

where we modeled the true distrib. $p(x)$ by an approx. one $q(x)$.

Since we optimize our code by $q(x)$,

we have to pay extra bandwidth to send the informations of outputs from p.

$KL(p \| q)$

- $KL(p \| q) > 0$ for $p \neq q$
- $KL(p \| q) = 0$ for $p = q$

- Mutual information

$$I[x,y] = KL\left(p(x,y) \,\middle|\, p(x)p(y)\right)$$

$$= -\int p(x,y) \ln \frac{p(x)p(y)}{p(x,y)}\, dx\, dy \qquad (\geq 0)$$

$$H[x] = H[x|y] + I[x,y]$$

PRML 2. Probability Distributions

§ 2.1 Binary Variables

- $\mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}$ $\quad [x = 0, 1]$ $\quad \mathbb{E} = \mu,\ \mathbb{V} = \mu(1-\mu)$
  - $1-\mu$: $x=0$, μ : $x=1$
- $\mathrm{Bin}(x|N,\mu) = {}_N C_x\, \mu^x (1-\mu)^{N-x}$ $\quad \mathbb{E} = N\mu,\ \mathbb{V} = N\mu(1-\mu)$
  - N: # trial, μ: mean
- $\mathrm{Beta}(x|a,b) = \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$ $\quad \mathbb{E} = \dfrac{a}{a+b},\ \mathbb{V} = \dfrac{ab}{(a+b)^2(a+b+1)}$
  - $x \in [0,1]$
  - a, b: hyperparameters
  - useful for prior
  - $\sim a^{-1}$ for $a = b$

Using $\mathrm{Beta}(x|a,b)$ as a prior, posterior after "$l \times m + 0 \times l$" obs. is

$$p(x|m,l,a,b) \propto \mu^{m+a-1}(1-\mu)^{l+b-1}$$
$$= \mathrm{Beta}(x|a+m, b+l)$$

"sequential learning"

- General discussion on Variance in sequential learning

where

$$\mathbb{E}_\theta[\theta] = \mathbb{E}_\mathcal{D}\left[\mathbb{E}_\theta[\theta|\mathcal{D}]\right] \tag{2.21}$$

- $\mathbb{E}_\theta[\theta]$: prior mean of a parameter
- $\mathbb{E}_\mathcal{D}[\cdot]$: averaged over data possibility
- $\mathbb{E}_\theta[\theta|\mathcal{D}]$: posterior mean of θ after $\mathcal{D}$-observation

$$\mathbb{E}_\theta[\theta] \equiv \int p(\theta)\theta \,\mathrm{d}\theta \tag{2.22}$$

$$\mathbb{E}_\mathcal{D}[\mathbb{E}_\theta[\theta|\mathcal{D}]] \equiv \int \left\{ \int \theta p(\theta|\mathcal{D}) \,\mathrm{d}\theta \right\} p(\mathcal{D}) \,\mathrm{d}\mathcal{D} \tag{2.23}$$

says that the posterior mean of $\boldsymbol{\theta}$, averaged over the distribution generating the data, is equal to the prior mean of $\boldsymbol{\theta}$. Similarly, we can show that

$$\mathrm{var}_\theta[\theta] = \mathbb{E}_\mathcal{D}\left[\mathrm{var}_\theta[\theta|\mathcal{D}]\right] + \mathrm{var}_\mathcal{D}\left[\mathbb{E}_\theta[\theta|\mathcal{D}]\right]. \tag{2.24}$$

prior variance = "averaged" posterior variance + (positive)

§ 2.2 Multinominal variables

Observation of "i" among "K" states $\Rightarrow$ $\mathbf{x} = \{0, \cdots, 0, \underset{\uparrow}{1}, 0, \cdots, 0\}$ (K; 1 at i-th element)

If the prob. to get "i" is μ_i,

$\sum \mu_i = 1$, $p(\mathbf{x}|\boldsymbol{\mu}) = \prod \mu_i^{x_i}$ $\cdots \sum x_i = 1$

$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = \{p(x_1|\mu), \cdots\} = \boldsymbol{\mu}^T$

- Likelihood to observe $\mathcal{D} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\mu}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \mu_k^{(x_n)_k} = \prod_k \mu_k^{\sum_n (x_n)_k}$$

$$= \prod_k \mu_k^{m_k} \quad m_k \Leftarrow \text{\# of observation of data } k$$

$\cdots$ "sufficient statistics"

$\Rightarrow (\mu_k)_{ML} = m_k/N$

- $\mathrm{Mult}(m_1, \cdots, m_K|\boldsymbol{\mu}, N) = \dfrac{N!}{m_1! \cdots m_N!} \prod_{k=1}^{K} \mu_k^{m_k}$ $\quad [\sum m_k = N]$

- $\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \dfrac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$

where $\alpha_0 = \sum \alpha_k$, $\sum \mu_k = 1$.

prior
$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \tag{2.38}$$

posterior
$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha} + \mathbf{m}) = \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1)\cdots\Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1} \tag{2.41}$$

§2.3 Gaussian Distribution

$$\mathcal{N}(\mathbf{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}\Delta^2\right]$$

- $\frac{D(D+3)}{2}$ params … too flexible
- Unimodal … too restricted

w. $\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$

- $\boldsymbol{\Sigma}$: Real Hermitian (symmetric) & positive definite
  $\Rightarrow$ · diagonalizable (eigenvalues $\in \mathbb{R}$)

$$\boldsymbol{\Sigma} = \mathbf{U}^\dagger\mathbf{L}\mathbf{U},\quad \Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^T\mathbf{U}\mathbf{L}^{-1}\mathbf{U}(\mathbf{x}-\boldsymbol{\mu}) =: \mathbf{y}^T\mathbf{L}^{-1}\mathbf{y}$$

$$\cdots\ \mathcal{N}(\mathbf{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{y_i^2}{2\lambda_i}\right]$$

$\mathbf{L} = \begin{pmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_N\end{pmatrix}$

- Writing $\mathcal{N}$ in terms of $\mathbf{y} = \mathbf{U}(\mathbf{x}-\boldsymbol{\mu})$

$$p(\mathbf{y}) = \left|\frac{d\mathbf{x}}{d\mathbf{y}}\right| p(\mathbf{x}) = |\mathbf{U}|^{-1}p(\mathbf{x}) = \prod_{i=1}^{D}\frac{1}{\sqrt{2\pi\lambda_i}}\exp\left(-\frac{y_i^2}{2\lambda_i}\right)$$

$\begin{cases}|\mathbf{U}| = 1\\ |\boldsymbol{\Sigma}| = |\mathbf{L}| = \prod\lambda_i\end{cases}$

- $\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu}$
- $\mathrm{Cov}[\mathbf{x}] = \boldsymbol{\Sigma}$

· for vector variables.

$$\mathrm{Cov}[\mathbf{x},\mathbf{y}] = \mathbb{E}\left((\mathbf{x}-\mathbb{E}(\mathbf{x}))(\mathbf{y}-\mathbb{E}(\mathbf{y}))^T\right) = \mathbb{E}(\mathbf{x}\mathbf{y}^T) - \mathbb{E}(\mathbf{x})\mathbb{E}(\mathbf{y}^T)$$

- Maximal Likelihood $p(\mathbf{X}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma})$

sufficient statistics $\sum \mathbf{x}_n$, $\sum \mathbf{x}_n\mathbf{x}_n^T$

$$\Rightarrow \boldsymbol{\mu}_{ML} = \frac{1}{N}\sum\mathbf{x}_n,\quad \boldsymbol{\Sigma}_{ML} = \frac{1}{N}\sum(\mathbf{x}_n-\boldsymbol{\mu}_{ML})(\mathbf{x}_n-\boldsymbol{\mu}_{ML})^T$$

(again, $\mathbb{E}(\boldsymbol{\mu}_{ML}) = \boldsymbol{\mu}$ but $\mathbb{E}(\boldsymbol{\Sigma}_{ML}) = \frac{N-1}{N}\boldsymbol{\Sigma}$.)

- $p(\mathbf{x}_a,\mathbf{x}_b)$ is Gaussian $\Longrightarrow$ $\begin{cases}p(\mathbf{x}_a\,|\,\mathbf{x}_b)\\ p(\mathbf{x}_a)\end{cases}$ are Gaussian.

§2.3.1 Conditional Gaussian distribution

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{y})}$$ ← y fixed, ← normalization

Separating $\mathbf{x}$ into $(\mathbf{x}_a, \mathbf{x}_b)$ and $\Lambda = \Sigma^{-1}$ into $\begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ab}^T & \Lambda_{bb} \end{pmatrix}$,

$$\Delta^2 = \begin{pmatrix} \mathbf{x}_a^T - \boldsymbol{\mu}_a^T & \mathbf{x}_b^T - \boldsymbol{\mu}_b^T \end{pmatrix} \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ab}^T & \Lambda_{bb} \end{pmatrix} \begin{pmatrix} \mathbf{x}_a - \boldsymbol{\mu}_a \\ \mathbf{x}_b - \boldsymbol{\mu}_b \end{pmatrix}$$

$$= (\mathbf{x}_a - \boldsymbol{\mu}_{a|b})^T \Lambda_{a|b} (\mathbf{x}_a - \boldsymbol{\mu}_{a|b}) + (\text{Remainders}) \quad \rightarrow \text{renormalized}$$

w.

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \Lambda_{aa}^{-1} \Lambda_{ab} (\mathbf{x}_b - \boldsymbol{\mu}_b)$$
$$= \boldsymbol{\mu}_a + \Sigma_{ab} \Sigma_{bb}^{-1} (\mathbf{x} - \boldsymbol{\mu}_b).$$

$$\Sigma_{a|b} = \Lambda_{aa}^{-1}$$
$$= \Sigma_{aa} - \Sigma_{ab} \Sigma_{bb}^{-1} \Sigma_{ab}^T$$

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix}$$

w. $M = (A - BD^{-1}C)^{-1}$

$$\begin{cases} \Lambda_{aa} = (\Sigma_{aa} - \Sigma_{ab} \Sigma_{bb}^{-1} \Sigma_{ab}^T)^{-1} \\ \Lambda_{ab} = -\Lambda_{aa} \Sigma_{ab} \Sigma_{bb}^{-1} \\ \Lambda_{bb} = \Sigma_{bb}^{-1} + \Sigma_{bb}^{-1} \Sigma_{ab}^T \Lambda_{aa} \Sigma_{ab} \Sigma_{bb}^{-1} \end{cases}$$

§2.3.2 Marginal Gaussian distribution

$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b$$

$$\mathbb{E}[\mathbf{x}_a] = \boldsymbol{\mu}_a$$
$$\text{Cov}[\mathbf{x}_a] = \Sigma_{aa}$$

Partitioned Gaussians

Given a joint Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ with $\Lambda \equiv \Sigma^{-1}$ and

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \tag{2.94}$$

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}. \tag{2.95}$$

Conditional distribution:

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{a|b}, \Lambda_{aa}^{-1}) \tag{2.96}$$
$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \Lambda_{aa}^{-1} \Lambda_{ab} (\mathbf{x}_b - \boldsymbol{\mu}_b). \tag{2.97}$$

Marginal distribution:

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \Sigma_{aa}). \tag{2.98}$$

§2.3.3 Bayes' theorem

$$\begin{cases} p(x) = \mathcal{N}(x \mid \mu, \Lambda^{-1}) \\ p(y \mid x) = \mathcal{N}(y \mid Ax + b, L^{-1}) \end{cases} \longrightarrow \text{construct} \begin{cases} p(x \mid y) \\ p(y) \end{cases}$$

linear in x ; indep. of x ; "linear Gaussian model"

Marginal and Conditional Gaussians

Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{2.113}$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{2.114}$$

the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \tag{2.115}$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \tag{2.116}$$

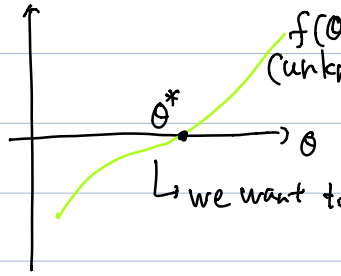
where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}. \tag{2.117}$$

(Straightforward w. $p(x) = p(x, y) = p(x)\,p(y \mid x)$

* § 2.3.5

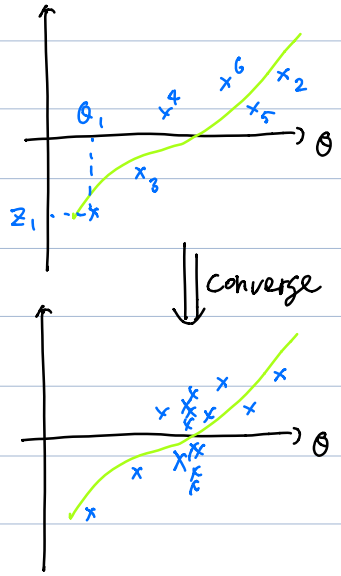
☆ Robbins-Monro algorithm



← must be:
Non-decreasing, and $f'(\theta^*) > 0$
(i.e. solution θ^* is unique)

↳ we want to find $f(\theta^*) = 0$.

has upper+lower bounds independent of θ.
must be uniformly bounded



• We have "measurements" z out of random variable $N(\theta)$
where $\mathbb{E}[N(\theta)] = f(\theta)$.
or: $\mathbb{E}[z|\theta]$

• A sequence $\{a_n\}$:

$$\sum_{n=0}^{\infty} a_n = \infty, \quad \sum_{n=0}^{\infty} a_n^2 < \infty, \quad \left(\text{and } \lim_{n\to\infty} a_n = 0\right)$$

$$\Rightarrow \theta_{n+1} = \theta_n - a_n \cdot \underbrace{z_n}_{\text{out of } N(\theta_n)} \xrightarrow[n\to\infty]{L^2\text{-converge}} \theta^*.$$

▣ Sequential ML Estimation

μ_{ML} maximizes $\prod_{n=1}^{N} p(x_n|\mu)$ i.e. solves $-\frac{1}{N}\frac{\partial}{\partial\mu}\sum_{n=1}^{N} \ln p(x_n|\mu) = 0$.

($N\to\infty$)

Thus the ideal value of μ_{ML} solves $\mathbb{E}\left[-\frac{\partial}{\partial\mu}\ln p(x|\mu)\right]\Big|_{\mu=\mu_{ML}} = 0$

$N(\theta)$

$$\therefore \mu_{ML}^{(N+1)} = \mu_{ML}^{(N)} - a_n \cdot \left(-\frac{\partial}{\partial\mu}\ln p(x_n|\mu)\Big|_{\mu=\mu_{ML}^{(N)}}\right)$$

will L^2-converge (almost sure converge) to $(\mu_{ML})^{ideal}$.

※ for Gaussian, $\mu_{ML}^{(N+1)} = \mu_{ML}^{(N)} + a_n \cdot \frac{x_n - \mu_{ML}^{(N)}}{\sigma^2}$

§2.3.6 Bayesian treatment

Remember: $p(\theta|x) \propto p(x|\theta)p(\theta)$ (posterior ∝ likelihood × prior)

- Unknown μ, known σ

$$p(X|\mu) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right]$$

⇒ A good prior: $\mathcal{N}(x|\mu_0,\sigma_0)$

posterior = PRIOR ⊕ data

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}} \quad (2.141)$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2}+\frac{N}{\sigma^2} \quad (2.142)$$

- Known μ, unknown $\sigma \equiv \lambda^{-1/2}$

$$p(X|\lambda) \propto \lambda^{N/2}\exp\left[-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right]$$

↳

$$\mathrm{Gam}(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}$$

$$\mathbb{E}[\lambda] = a/b, \quad \mathbb{V}[\lambda] = a/b^2$$

$$a_N = a_0 + \frac{N}{2} \quad (2.150)$$

$$b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 = b_0 + \frac{N}{2}\sigma_{\mathrm{ML}}^2 \quad (2.151)$$

- Both Unknown

$$p(X|\mu,\lambda) \propto \lambda^{N/2}\exp\left[-\frac{\lambda}{2}\Sigma(x-\mu)^2\right]$$

$$= \lambda^{N/2}\exp\left[-\frac{N\lambda}{2}\left((\mu-\mu_{ML})^2+\sigma_{ML}^2\right)\right]$$

$$\propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right\}. \quad (2.152)$$

↳ Gaussian-gamma distribution

$$GG(\mu,\lambda|\mu_0,\beta,a,b) = \mathcal{N}(\mu|\mu_0,(\beta\lambda)^{-1})\,\mathrm{Gam}(\lambda|a,b)$$

※ Multivariable case

μ - unknown: Gaussian

σ - unknown: Wishart

$$\mathcal{W}(\mathbf{\Lambda}|\mathbf{W},\nu) = B|\mathbf{\Lambda}|^{(\nu-D-1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\mathbf{\Lambda})\right) \quad (2.155)$$

both unknown:

$$p(\boldsymbol{\mu},\mathbf{\Lambda}|\boldsymbol{\mu}_0,\beta,\mathbf{W},\nu) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0,(\beta\mathbf{\Lambda})^{-1})\,\mathcal{W}(\mathbf{\Lambda}|\mathbf{W},\nu) \quad (2.157)$$

where

$$B(\mathbf{W},\nu) = |\mathbf{W}|^{-\nu/2}\left(2^{\nu D/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1}. \quad (2.156)$$

§2.3.7 Student's t-distribution : infinite mixture of Gaussian

$$St(x|\mu,\lambda,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]^{-(\nu+1)/2}$$

this appears in the marginal distrib.

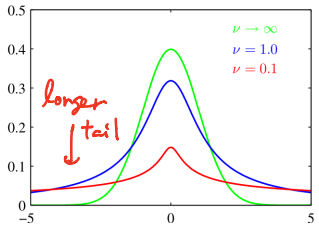
$$p(\mu|\mu_0,\beta=1,a,b) = \int_0^\infty GG(\mu,\lambda|\mu_0,1,a,b)\,d\lambda$$

$$= \int_0^\infty \mathcal{N}(\mu|\mu_0,\lambda^{-1})\,Gam(\lambda|a,b)\,d\lambda$$

$$= \frac{\Gamma(a+\frac{1}{2})}{\Gamma(a)}\left(\frac{1}{2\pi b}\right)^{1/2}\left[1+\frac{(\mu-\mu_0)^2}{2b}\right]^{-\frac{2a+1}{2}}$$

$$St(x|\mu,\lambda,\nu\to\infty) = \mathcal{N}(x|\mu,\lambda^{-1})$$

$$St(x|\mu,\lambda,\nu=1) = \text{Cauchy distribution}$$



↳ robustness

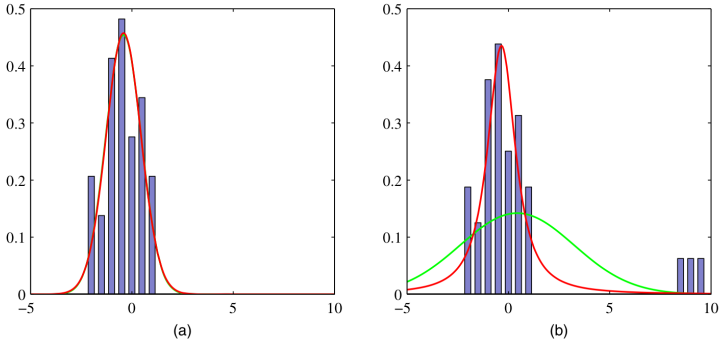


Figure 2.16 Illustration of the robustness of Student's t-distribution compared to a Gaussian. (a) Histogram distribution of 30 data points drawn from a Gaussian distribution, together with the maximum likelihood fit obtained from a t-distribution (red curve) and a Gaussian (green curve, largely hidden by the red curve). Because the t-distribution contains the Gaussian as a special case it gives almost the same solution as the Gaussian. (b) The same data set but with three additional outlying data points showing how the Gaussian (green curve) is strongly distorted by the outliers, whereas the t-distribution (red curve) is relatively unaffected.

* Multivariate version

$$\mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) = \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1})\mathrm{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\eta. \quad (2.161)$$

$$= \frac{\Gamma(D/2+\nu/2)}{\Gamma(\nu/2)}\frac{|\boldsymbol{\Lambda}|^{1/2}}{(\pi\nu)^{D/2}}\left[1+\frac{\Delta^2}{\nu}\right]^{-D/2-\nu/2} \quad (2.162)$$

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})$$

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}, \quad \text{if} \quad \nu>1 \quad (2.164)$$

$$\mathrm{cov}[\mathbf{x}] = \frac{\nu}{(\nu-2)}\boldsymbol{\Lambda}^{-1}, \quad \text{if} \quad \nu>2 \quad (2.165)$$

$$\mathrm{mode}[\mathbf{x}] = \boldsymbol{\mu} \quad (2.166)$$

§2.3.8 Periodic variable : von Mises distribution

$$p(\theta \mid \theta_0, m) = \frac{1}{2\pi I_0(m)} e^{m\cos(\theta-\theta_0)}$$

θ: $[0, 2\pi]$ periodic

where $I_0(m) = \frac{1}{2\pi}\int_0^{2\pi} \exp\{m\cos\theta\}\, d\theta.$

this corresponds to $N\left(\begin{pmatrix}x\\y\end{pmatrix} \middle| \begin{pmatrix}x_0\\y_0\end{pmatrix}, \begin{pmatrix}\sigma & 0\\0 & \sigma\end{pmatrix}\right)$ restricted on $x^2+y^2=1$.

$$\begin{cases} \theta_0^{ML} = \text{Arctan}\dfrac{\Sigma \sin\theta_n}{\Sigma\cos\theta_n} \\ \dfrac{I_1(m_{ML})}{I_0(m_{ML})} = \dfrac{1}{N}\Sigma\cos(\theta_n - \theta_{ML}) \end{cases}$$

$$= \left(\frac{1}{N}\Sigma\cos\theta_n\right)\cos\theta_0^{ML} + \left(\frac{1}{N}\Sigma\sin\theta_n\right)\sin\theta_0^{ML}$$

§2.3.9 Mixtures of Gaussians

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k) \quad \text{w.} \begin{cases} \pi_k \in [0,1] \\ \Sigma \pi_k = 1. \end{cases}$$

We can view this as $p(x) = \sum_k p(k)\, p(x \mid k)$ (prior to choose k)

Then

$$p(k \mid x) = \frac{p(k)p(x\mid k)}{\Sigma p(i) p(x\mid i)} = \frac{\pi_k \mathcal{N}(x\mid \mu_k, \Sigma_k)}{\Sigma \pi_i \mathcal{N}(x\mid \mu_i, \Sigma_i)}$$

$\equiv \gamma_k(x)$: responsibilities

§2.4 Exponential Family

$$p(x|\eta) = g(\eta) \cdot h(x) e^{\eta^T \cdot u(x)}$$

normalization

$$\left(\equiv \exp\left[\eta^T \cdot u(x) + \tilde{g}(\eta) + \tilde{h}(x) \right] \right)$$

		$h(x)$	η	u
$\text{Bern}(x|\mu) \propto e^{x \log \frac{\mu}{1-\mu}}$	:	1	$\ln \frac{\mu}{1-\mu}$	x
$\text{Beta}(x|a,b) \propto \frac{1}{x(1-x)} e^{a \log x + b \log(1-x)}$	:	$\frac{1}{x(1-x)}$	a	$\log x$
			b	$\log(1-x)$
$\text{Bin}(x|N,\mu) \propto {}_N C_x \, e^{x \log \frac{\mu}{1-\mu}}$ (FIXED)	:	${}_N C_x$	$\ln \frac{\mu}{1-\mu}$	x
$\mathcal{N}(x|\mu,\lambda^{-1/2}) \propto e^{-\frac{\lambda}{2}(x^2 - 2x\mu)}$	:	1	$-\lambda/2$	x^2
			$\lambda\mu$	x
$\text{Pois}(n|\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$	:	$\frac{1}{n!}$	$\ln \lambda$	n

$$\text{Mult}(m_1, \cdots, m_N \,|\, \mu, N) = \frac{N!}{m_1! \cdots m_N!} \exp\left[\sum m_k \log \mu_k \right]$$

(FIXED)

$$= \quad " \quad \exp\left[\sum_{k=1}^{N-1} m_k \log \mu_k + \left(N - m_1 - \cdots \right) \log\left(1 - \mu_1 - \cdots \right) \right]$$

$$\therefore h(x) = \frac{N!}{m_1! \cdots m_{N-1}! \left(1 - \sum_{k=1}^{N-1} m_k \right)!}$$

$$u = \begin{pmatrix} m_1 \\ \vdots \\ m_{N-1} \end{pmatrix}$$

$$\eta_k = \log \frac{\mu_k}{1 - \sum_{i=1}^{N-1} \mu_i}$$

$$\left(\Leftrightarrow \mu_k = \frac{\exp \eta_k}{1 + \sum_{i=1}^{N-1} \exp \eta_i} \right)$$

- Formulae $-\nabla_\eta \ln g(\eta) = \mathbb{E}[u(x)]$

- ML method

η_{ML} solves $0 = \nabla_\eta \left(p(X|\eta) \right)$

$$= \nabla_\eta \left[\prod_{n=1}^{N} h(x_n) \cdot g(\eta)^N \cdot \exp\left[\eta^T \cdot \sum_{n=1}^{N} u(x_n) \right] \right]$$

$$\therefore \quad \boxed{\nabla_\eta \ln g(\eta)\Big|_{\eta=\eta_{ML}} = \frac{1}{N} \sum_{n=1}^{N} u(x_n)}$$

sufficient statistics

$\Sigma u \to N\bar{x}$, $N \to \nu$

- Conjugate prior

$$p(\eta | \chi, \nu) = f(\chi, \nu)\, g(\eta)^\nu \exp\left[\nu \eta^T \chi \right]$$

↓ posterior

$$p(\eta | X, \chi, \nu) = \tilde{f}(X, \chi, \nu)\, g(\eta)^{\nu+N} \exp\left[\eta^T \left(\nu\chi + \sum u(x_n) \right) \right]$$

ⓓ Non-informative prior

- Constant prior? … ① improper if unbound

② constant on which parametrization? … take care!

- translation-invariant prior

… inv. under $x \to x + c$. has form $p(x|\mu) = f(x-\mu)$.

… $p(\mu) = \text{const.} \equiv \lim_{\sigma_0 \to \infty} N(\mu | \mu_0, \sigma_0)$ — μ: location parameter

- scale-invariant prior

… inv. under $x \to k x$; form of $p(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$ — σ: scale parameter

… $p(\sigma) \propto 1/\sigma \equiv \lim_{\substack{a \to 0 \\ b \to 0}} \text{Gam}(\sigma | a, b)$

§2.5 Nonparametric methods

$$p(x) = \frac{K}{N V}$$

K ← # "near" events; N = #events; V ← volume of "near"

- **Histogram**
  - count events within bins.

  [careful on the bin-width! very bad for large dim. data]

- **Kernel density estimators**
  - count events that fall nearby.

$$p(x) = \frac{1}{N} \cdot \frac{1}{h^D} \sum_{n=1}^{N} k\left(\frac{x - x_n}{h}\right)$$

x_n: location of n-th event

where $k(d) = \begin{cases} 1 & : \forall i,\ d_i \le \frac{1}{2} \\ 0 & : \text{otherwise} \end{cases}$

OR

$$p(x) = \frac{1}{N} \frac{1}{(2\pi h^2)^{D/2}} \sum_{n=1}^{N} \exp\left[-\frac{\|x - x_n\|^2}{2h^2}\right]$$

$k(d)$: $k(d) \ge 0$ and $\int k(d)\, d\, d = 1$

- **K-nearest neighbor method**
  - calculate sphere volume V that includes K neighbors.

needs entire data $\{x_n\}$ for evaluation !!

Histogram | Kernel density | K-nearest neighbor