Resumo de Aprendizagem de Máquina 2014-2

Eduardo M. B. de A. Tenório
embat@cin.ufpe.br

CIn-UFPE

Resumo

Este documento tem por finalidade ser um resumo dos assuntos abordados na disciplina Aprendizagem de Máquina *do período 2014-2 do CIn-UFPE, ministrada pelos professores Francisco Carvalho e Teresa Ludermir. A maioria do documento referencia o livro "Pattern Classification", de Duda, Hart & Stork. Os códigos utilizados como exercício de fixação encontram-se em* github.com/embatbr/resumo-aprendizagem.

1 Teoria da Decisão Bayesiana

1.1 Introdução

Teoria da Decisão Bayesiana é uma abordagem estatística para a classificação de padrões, baseada em quantificar os tradeoffs associados a tomar uma determinada decisão (classificar) utilizando probabilidade e considerando os custos associados.

O **estado natural** é denotado por ω, de modo que $\omega = \omega_i$, para $i = 1, 2, ..., c$, significa que o exemplo foi classificado como pertencente à classe ω_i. Cada uma dessas classes possui uma **probabilidade a priori** $P(\omega_i)$, com

$$\sum_{i=1}^{c} P(\omega_i) = 1, \qquad (1)$$

refletindo o conhecimento prévio da chance de um elemento da classe ω_i aparecer. A **regra de decisão** fica:

$$\text{Decida } \omega_i \text{ se } i = \arg\max_j P(\omega_j). \qquad (2)$$

Neste caso a classe ω_i sempre é escolhida e a probabilidade de erro é dada por:

$$P_{err}(\omega_i) = 1 - P(\omega_i). \qquad (3)$$

Utilizando uma característica x que seja contínua e aleatória, sua **densidade de probabilidade estado-condicional** é dada por $p(x|\omega)$. Logo, a diferença entre $p(x|\omega_i)$ e $p(x|\omega_j)$ descreve a diferença da característica x entre as populações das classes ω_i e ω_j.

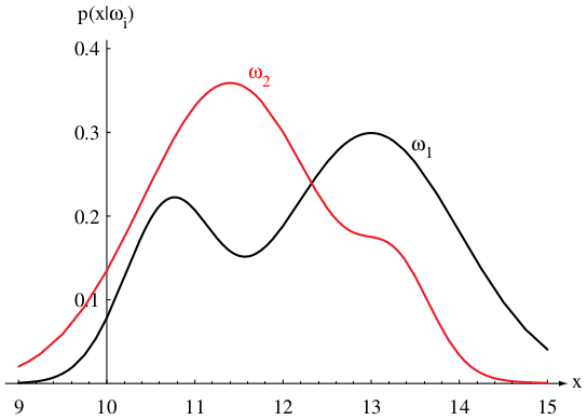

Figura 1: Para $\omega_i = \omega_2$, é mais frequente observar x entre 11 e 12 que $x = 13$ (valor mais provável se $\omega_i = \omega_1$).

Sabendo $P(\omega_i)$ e $p(x|\omega_i)$, e medindo um valor x, a probabilidade conjunta de achar

um padrão na classe ω_i e com x é dado por: $p(\omega_i, x) = P(\omega_i|x)p(x) = p(x|\omega_i)P(\omega_i)$, que pela **fórmula de Bayes** fica:

$$P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{p(x)}, \quad (4)$$

com a evidência para c classes

$$p(x) = \sum_{j=1}^{c} p(x|\omega_j)P(\omega_j). \quad (5)$$

A probabilidade a posteriori das classes ω_1 e ω_2 para um conjunto de valores de x é mostrada em Fig. (2). A regra de decisão fica:

$$\text{Decida } \omega_i \text{ se } i = \arg\min_j P(erro_j|x), \quad (6)$$

onde

$$P(erro_i|x) = \sum_{j \neq i} P(\omega_j|x), \quad (7)$$

ou simplesmente $P(erro_i|x) = 1 - P(\omega_i|x)$. Então a regra torna-se:

$$\text{Decida } \omega_i \text{ se } i = \arg\max_j P(\omega_j|x), \quad (8)$$

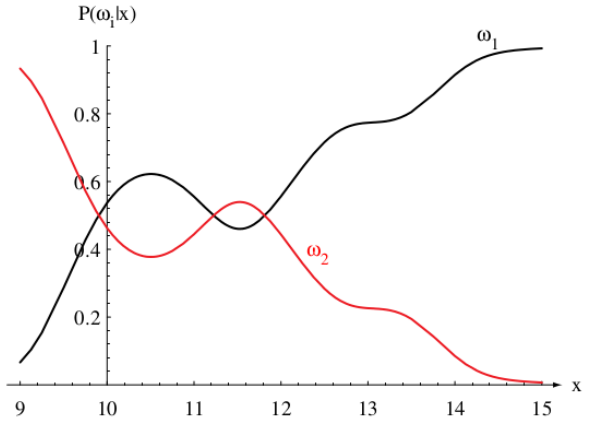

Figura 2: Probabilidades a posteriori para $P(\omega_1) = \frac{2}{3}$ e $P(\omega_2) = \frac{1}{3}$, e para as densidades de probabilidade estado-condicional mostradas em Fig. (1).

Esta regra minimiza a probabilidade média de erro, dada por

$$P(erro_i) = \int_{-\infty}^{\infty} P(erro_i|x)p(x)dx. \quad (9)$$

1.2 Características Contínuas

É de fácil compreensão que a característica x pode ser trocada por um vetor de características $\boldsymbol{x} = (x_1, x_2, ..., x_d)$, onde $\boldsymbol{x}$ pertence ao espaço $\boldsymbol{R}^d$ (espaço de características). A região que decide ω_i é denotada por $\mathcal{R}_i$.

Outras ações além de apenas classificar um elemento podem ser tomadas, como por exemplo a **rejeição**: recusar-se a tomar uma decisão; uma opção válida quando o custo de ser indeciso é aceitável. Para isso **funções de custo** são inseridas, permitindo tratar de situações onde alguns erros de classificação são mais importantes que outros.

Seja $\{\omega_1, ..., \omega_c\}$ o conjunto finito de c classes e seja $\{\alpha_1, ..., \alpha_a\}$ o conjunto finito de possíveis ações. A função de custo $\lambda(\alpha_i|\omega_j)$ descreve o custo de tomar a ação α_i quando a classe é ω_j. Logo, observado um $\boldsymbol{x}$ em particular, tomar a ação α_i quando a classe é ω_j leva a um custo esperado (**risco**)

$$R(\alpha_i|\boldsymbol{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\boldsymbol{x}). \quad (10)$$

$R(\alpha_i|\boldsymbol{x})$ é chamado de **risco condicional**. Qualquer que seja o $\boldsymbol{x}$ observado, o risco pode ser minimizado selecionando a ação que minimiza $R(\alpha_i|\boldsymbol{x})$.

A regra de decisão geral é uma função $\alpha(\boldsymbol{x})$ que diz qual ação tomar para cada possível observação, ou seja, para cada $\boldsymbol{x}$ a **função de decisão** $\alpha(\boldsymbol{x})$ assume um dos a valores $\alpha_1, ..., \alpha_a$.

$$\alpha(\boldsymbol{x}) = \alpha_i \text{ se } i = \arg\min_j R(\alpha_j|\boldsymbol{x}), \quad (11)$$

Logo, o **risco global** é dado por

$$R = \int R(\alpha(\boldsymbol{x}))p(\boldsymbol{x})d\boldsymbol{x}. \quad (12)$$

O risco global mínimo é chamado de **risco de Bayes**, denotado por R^*, sendo a melhor perfomance alcançável.

1.3 Taxa de Erro Mínima

Para evitar erros, a regra de decisão procurada é aquela que minimiza a probabilidade de erro, i.e. minimiza a **taxa de erro**. A função de custo de interesse para este caso é chamada de **simétrica** ou **zero-um**,

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & \text{se } i = j \\ 1 & \text{se } i \neq j \end{cases} \qquad (13)$$

para $i, j = 1, ..., c$. Como todos os erros tem custo igual, o risco condicional é dado por

$$R(\alpha_i|\boldsymbol{x}) = 1 - P(\omega_i|\boldsymbol{x}) \qquad (14)$$

com $P(\omega_i|\boldsymbol{x})$ sendo a probabilidade condicional da ação α_i estar correta. A regra de decisao neste caso continua:

$$\text{Decida } \omega_i \text{ se } i = \arg\max_j P(\omega_j|x). \qquad (15)$$

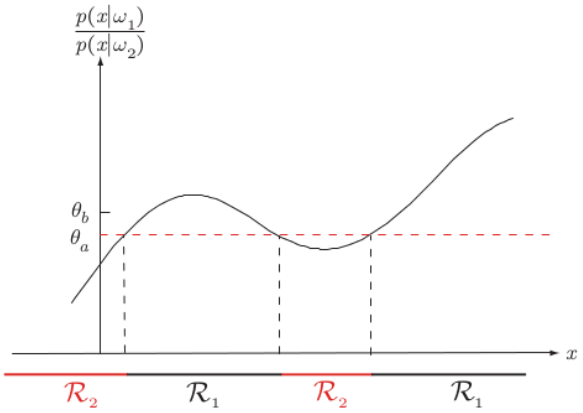

Figura 3: Se a penalização de classificar ω_1 como ω_2 for maior que o oposto, então a razão tende ao threshold θ_b.

1.4 Funções Discriminantes

A maneira mais usual de representar classificadores de padrões é através de um conjunto de **funções discriminantes** $g_i(\boldsymbol{x}), i = 1, ..., c$. O classificador atribui um vetor de características $\boldsymbol{x}$ à classe ω_i se

$$g_i(\boldsymbol{x}) = \arg\max_j g_j(\boldsymbol{x}) \qquad (16)$$

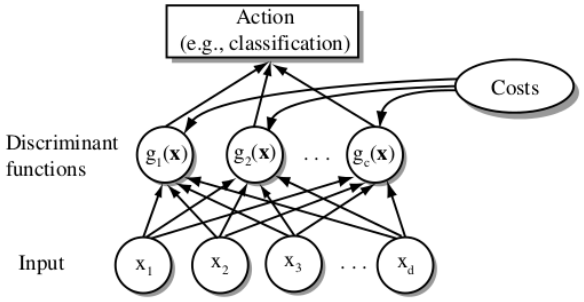

Figura 4: Classificador com c funções discriminantes e entradas d-dimensional. A ação geralmente é "escolher o maior $g_i(\boldsymbol{x})$".

Para o caso geral com riscos, pode-se fazer $g_i(\boldsymbol{x}) = -R(\alpha_i|\boldsymbol{x})$, enquanto para o caso "taxa de erro mínima", $g_i(\boldsymbol{x}) = P(\omega_i|\boldsymbol{x})$. A função discriminante $g_i(\boldsymbol{x})$ pode ser substituída por $f(g_i(\boldsymbol{x}))$, com $f(\cdot)$ sendo uma função monotonicamente crescente (e.g. logaritmo), com o resultado da classificação ficando inalterado. Como resultado, $\boldsymbol{R}^d$ é dividido em regiões de decisão $\mathcal{R}_i$ (não necessariamente conectadas) para cada classe ω_i.

Para o caso em que $c = 2$, o classificador é chamado **dicotomizador**, e apenas uma função discriminante $g(\boldsymbol{x}) \equiv g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$ é necessária. Logo a regra de decisão torna-se:

$$\text{Decida } \omega_1 \text{ se } g(\boldsymbol{x}) > 0; \text{ senão, } \omega_2. \qquad (17)$$

1.5 Características Discretas

Em muitas aplicações práticas as componentes de $\boldsymbol{x}$ são valores inteiros binários, ternários ou outro "ário", de modo que $\boldsymbol{x}$ pode assumir apenas um dos m valores discretos $\boldsymbol{v_1}, ..., \boldsymbol{v_m}$. Nestes casos, a função de densidade de probabilidade $p(\boldsymbol{x}|\omega_j)$ torna-se uma função de massa de probabilidade $P(\boldsymbol{x}|\omega_j)$ e

$$\int p(\boldsymbol{x}|\omega_j)d\boldsymbol{x} \qquad (18)$$

é substituída por

$$\sum_{\boldsymbol{x}} P(\boldsymbol{x}|\omega_j). \qquad (19)$$

Na fórmula de Bayes as densidades de probabilidade são trocadas por probabilidades

$$P(\omega_j|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\omega_j)P(\omega_j)}{P(\boldsymbol{x})} \qquad (20)$$

onde

$$P(\boldsymbol{x}) = \sum_{j=1}^{c} P(\boldsymbol{x}|\omega_j)P(\omega_j). \qquad (21)$$

A definição do risco condicional $R(\alpha|\boldsymbol{x})$ mantém-se inalterada.

2 Estimação Paramétrica

2.1 Introdução

Sabendo $P(\omega_i)$ e $p(\boldsymbol{x}|\omega_i)$ é possível projetar um classificador ótimo. Contudo, em aplicações reais de classificação de padrões raramente tem-se este conhecimento completo a respeito da estrutura probabilística do problema. Geralmente o que está disponível é um conhecimento superficial da situação e um conjunto de **dados de treinamento**.

Como exemplo, sejam 300 objetos divididos em 2 classes de modo que a classe 1 tenha 200 objetos e a classe 2 tenha 100. Cada objeto é gerado a partir de uma das 3 gaussianas bi-variadas:

1a: $\mu_1 = 60$, $\mu_2 = 30$, $\sigma_1^2 = 9$ e $\sigma_2^2 = 144$

1b: $\mu_1 = 52$, $\mu_2 = 30$, $\sigma_1^2 = 9$ e $\sigma_2^2 = 9$

2: $\mu_1 = 45$, $\mu_2 = 22$, $\sigma_1^2 = 100$ e $\sigma_2^2 = 9$

Nota-se que $\boldsymbol{\Sigma}_{1a}$, $\boldsymbol{\Sigma}_{1b}$ e $\boldsymbol{\Sigma}_2$ são matrizes diagonais, reduzindo-as a vetores de variância, o que significa independência entre as variáveis.

As duas imagems em Fig. (5) mostram os dados de treinamento. Na de cima os dados são divididos em clusters gerados pelas gaussianas definidas anteriormente. Na segunda são dividos em duas classes.

Pode-se utilizar estas amostras para estimar as probabilidades e densidades de probabilidade desconhecidas. Para problemas

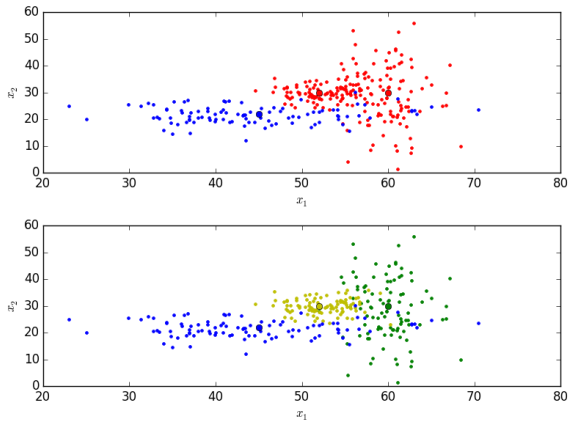

Figura 5: Classes 1a (verde), 1b (amarelo) e 2 (azul).

de aprendizagem supervisionada, a estimação das $P(\omega_i)$ é bastante simples. Sabendo a frequência relativa de aparição de cada classe ω_i no conjunto de treinamento, pode-se estimar suas $P(\omega_i)$ (desde que o conjunto de treinamento seja uma boa representação do "mundo real"). Contudo, estimar as $p(\boldsymbol{x}|\omega_i)$ é complicado quando se tem poucos dados, especialmente quando a dimensionalidade de $\boldsymbol{x}$ aumenta. Uma maneira mais prática é assumir cada $p(\boldsymbol{x}|\omega_i)$ como uma gaussiana, cujas média e matriz de covariância são dadas por $\boldsymbol{\mu}_i$ e $\boldsymbol{\Sigma}_i$. Logo, o problema passa de estimar uma função desconhecida $p(\boldsymbol{x}|\omega_i)$ para estimar os **parâmetros** $\boldsymbol{\mu}_i$ e $\boldsymbol{\Sigma}_i$.

Para a estimação dos parâmetros, será utilizada a técnica da **máxima-verossimilhança**.

2.2 Máxima-Verossimilhança

Estimação por Máxima-Verossimilhança (Maximum-Likelihood, ML) considera os parâmetros como quantidades cujos valores são fixos, mas desconhecidos. A melhor estimativa de seus valores é aquela que maximiza a probabilidade de obter as amostras observadas. Este método possui uma gama de atributos atrativos. Primeiramente, possui quase sempre boas propriedades de convergência quando o número de amostras de treinamento aumenta. Além disso, ML

geralmente é mais simples que métodos alternativos, tais como técnicas bayesianas.

Princípio Geral

Um conjunto $\mathcal{D}$ de amostras é dividido em conjuntos $\mathcal{D}_1, .., \mathcal{D}_c$ cujas amostras são desenhadas independentemente, de acordo com $p(\boldsymbol{x}|\omega_j)$, resultado de variáveis aleatórias independentes e identicamente distribuídas. Assume-se que $p(\boldsymbol{x}|\omega_j)$ tem uma forma paramétrica, como por exemplo $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, de modo que pode ser escrita como $p(\boldsymbol{x}|\omega_j, \boldsymbol{\theta}_j)$, onde $\boldsymbol{\theta}_j$ consiste nos componentes de $\boldsymbol{\mu}_j$ e $\boldsymbol{\Sigma}_j$. O problema torna-se utilizar as informações dadas pelas amostras para obter boas estimativas dos vetores desconhecidos $\boldsymbol{\theta}_j$, para cada ω_j. Uma simplificação é assumir que as amostras de cada $\mathcal{D}_j$ provêem informações relevantes apenas aos seus respectivos $\boldsymbol{\theta}_j$.

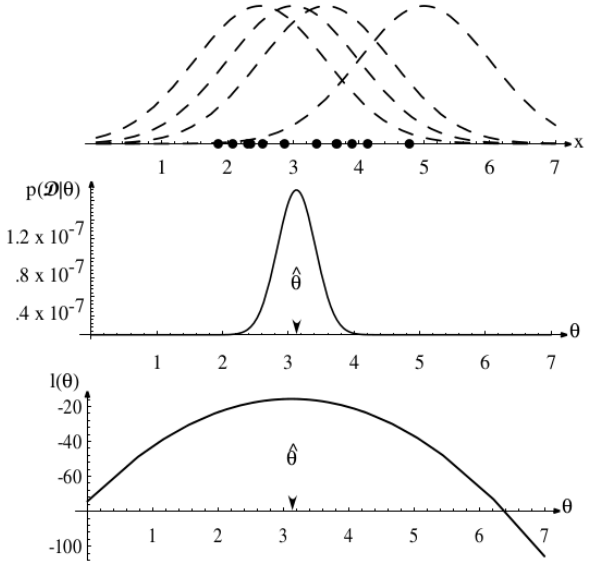

Figura 6: Estimação do parâmetro $\hat{\boldsymbol{\theta}}$, o valor $\hat{\boldsymbol{\theta}}$ achado e $l(\hat{\boldsymbol{\theta}})$.

Usa-se o conjunto $\mathcal{D}$, composto pelas amostras de treinamento $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ desenhadas independentemente de $p(\boldsymbol{x}|\boldsymbol{\theta})$, para estimar os $\boldsymbol{\theta}$ desconhecidos, tal que

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\boldsymbol{x}_k|\boldsymbol{\theta}). \qquad (22)$$

A densidade $p(\mathcal{D}|\boldsymbol{\theta})$ é conhecida como a **verossimilhança** de $\boldsymbol{\theta}$ em relação às amostras, e sua estimativa máxima é, por definição, o valor $\hat{\boldsymbol{\theta}}$ que maximiza $p(\mathcal{D}|\boldsymbol{\theta})$ (Fig. (6)).

Devido a ser monotonicamente crescente e mais fácil de trabalhar que a verossimilhança, o uso da log-verossimilhança

$$l(\boldsymbol{\theta}) \equiv \ln p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{k=1}^{n} \ln p(\boldsymbol{x}_k|\boldsymbol{\theta}) \qquad (23)$$

é preferível. Logo, a solução para $\hat{\boldsymbol{\theta}}$ é

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}), \qquad (24)$$

onde a dependencia do conjunto de dados $\mathcal{D}$ fica implícito. Se o número de parametros a serem estimados é p, então $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^t$ e $\boldsymbol{\nabla}_{\boldsymbol{\theta}} = (\frac{\partial}{\partial\theta_1}, ..., \frac{\partial}{\partial\theta_p})^t$, e, pela Eq. (23)

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} l = \sum_{k=1}^{n} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p(\boldsymbol{x}_k|\boldsymbol{\theta}). \qquad (25)$$

Então, um conjunto de condições necessárias para a estimação da máxima verossimilhança para $\boldsymbol{\theta}$ pode ser obtido do conjunto de p equações

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} l = \mathbf{0}. \qquad (26)$$

Uma solução $\hat{\boldsymbol{\theta}}$ para Eq. (26) pode ser uma máximo global, um máximo/mínimo local, ou (raramente) um ponto de inflexão de $l(\boldsymbol{\theta})$. Se todas as soluções forem achadas, é garantido que representa o máximo verdadeiro, senão deve-se checar todas as soluções individualmente (ou calcular derivadas de segunda ordem) para identificar qual é o ótimo global.

Gaussiana com μ desconhecido

Neste caso o vetor $\boldsymbol{\theta} = \boldsymbol{\mu}$, levando a $l(\boldsymbol{\theta}) = l(\boldsymbol{\mu}) \equiv \ln p(\mathcal{D}|\boldsymbol{\mu})$. Logo pela Eq. (23),

$$l(\boldsymbol{\mu}) = \sum_{k=1}^{n} \ln p(\boldsymbol{x}_k|\boldsymbol{\mu}), \tag{27}$$

e pela Eq. (25),

$$\boldsymbol{\nabla}_{\boldsymbol{\mu}} l = \sum_{k=1}^{n} \boldsymbol{\nabla}_{\boldsymbol{\mu}} \ln p(\boldsymbol{x}_k|\boldsymbol{\mu}). \tag{28}$$

Como $\boldsymbol{\nabla}_{\boldsymbol{\mu}} \ln p(\boldsymbol{x}_k|\boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu})$ e $\boldsymbol{\nabla}_{\boldsymbol{\mu}} l = 0$ (Eq. (26)), então

$$\sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}) = \mathbf{0}. \tag{29}$$

Multiplicando por $\boldsymbol{\Sigma}$ e rearranjando os termos, tem-se

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{x}_k. \tag{30}$$

Fica evidente que para o caso em que apenas $\boldsymbol{\mu}$ é desconhecido, a estimativa para a máxima verossimilhança é apenas a média das amostras, às vezes escrita como $\hat{\boldsymbol{\mu}}_n$ para clarificar sua dependência do número de amostras.

Gaussiana com μ e Σ desconhecidos

Este é o caso mais típico, quando $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ são desconhecidos. Desta forma, este parâmetros constituem as componentes do vetor paramétrico $\boldsymbol{\theta}$, ou seja $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})^t$. Considerando o caso univariado, $\theta_1 = \mu$ e $\theta_2 = \sigma^2$. Então

$$\ln p(x_k|\boldsymbol{\theta}) = -\frac{1}{2} \ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2 \tag{31}$$

e seu gradiente é

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} l = \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p(x_k|\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix} \tag{32}$$

e pela Eq. (29) tem-se

$$\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 \tag{33}$$

e

$$-\sum_{k=1}^{n} \frac{1}{2\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\theta}_1)^2}{2\hat{\theta}_2^{\,2}} = 0, \tag{34}$$

Rearrumando tudo, tem-se

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^{n} x_k \tag{35}$$

e

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \hat{\mu})^2. \tag{36}$$

Dando um salto de fé (embora esta seja demonstrável, diferente da cristã) e trocando os $\hat{\mu}$ e $\hat{\sigma}$ por $\hat{\boldsymbol{\mu}}$ e $\hat{\boldsymbol{\Sigma}}$, respectivamente, tem-se

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{x}_k \tag{37}$$

e

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{k=1}^{n} (\boldsymbol{x}_k - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}})^t. \tag{38}$$

3 Misturas

3.1 Introdução

Esta seção trata sobre aprendizado **não-supervisionado**, no qual as amostras não são rotuladas com suas classes. Existem pelo menos cinco razões básicas para utilizá-lo: (1) coletar e rotular um grande conjunto de dados

pode ser extremamente custoso; (2) o procedimento pode ser efetuado na ordem inversa, treinando com um grande número de amostras não rotuladas e então utilizar supervisão para rotular os grupos criados; (3) em muitas aplicações as características dos padrões podem mudar com o tempo; (4) métodos não-supervisionados podem ser utilizados para encontrar características que serão úteis para a categorização; (5) em estágio iniciais de uma investigação, pode ser interessante efetuar análise exploratória de dados para ganhar algum *insight* sobre a natureza ou estrutura dos dados.