University of Rome "La Sapienza"

Master in Artificial Intelligence and Robotics

Machine Learning

A.Y. 2018/2019

Prof. Luca Iocchi

Sapienza University of Rome, Italy
Master in Artificial Intelligence and Robotics
Machine Learning (2018/19)

7. Probabilistic models for classification

Luca Iocchi

Overview

- Probabilistic generative models
- Probabilistic discriminative models
- Logistic regression

References

C. Bishop. Pattern Recognition and Machine Learning. Sect. 4.2, 4.3

Probabilistic Models for Classification

- Generative: estimate $P(C_i|\mathbf{x})$ through $P(\mathbf{x}|C_i)$ and Bayes theorem
- Discriminative: estimate $P(C_i|\mathbf{x})$ directly from a model

Probabilistic Generative Models

Consider first the case of two classes.

Find the conditional probability:

$$
\begin{aligned}
P(C_1|\mathbf{x}) &= \frac{p(\mathbf{x}|C_1)P(C_1)}{p(\mathbf{x}|C_1)P(C_1) + p(\mathbf{x}|C_2)P(C_2)} \\
&= \frac{1}{1+\exp(-\alpha)} = \sigma(\alpha)
\end{aligned}
$$

with:

$$\alpha = \ln \frac{p(\mathbf{x}|C_1)P(C_1)}{p(\mathbf{x}|C_2)P(C_2)}$$

and

$\sigma(\alpha) = \frac{1}{1+\exp(-\alpha)}$ the *sigmoid function*.

Sapienza University of Rome, Italy - Machine Learning (2018/2019)

Probabilistic Generative Models

Assume $p(\mathbf{x}|C_i) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ - same covariance matrix

we get:

$$P(C_1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + w_0),$$

with:

$$
\begin{aligned}
\mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \\
w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{P(C_1)}{P(C_2)}.
\end{aligned}
$$

Probabilistic Generative Models

Maximum likelihhod solution for 2 classes

Assuming $P(C_1) = \pi$ (thus $P(C_2) = 1 - \pi$), $P(\mathbf{x}|C_i) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$

Given data set $D = \{(\mathbf{x}_n, t_n)_{n=1}^N\}$, $t_n = 1$ if $\mathbf{x}_n$ belongs to class C_1, $t_n = 0$ if $\mathbf{x}_n$ belongs to class C_2

Let N_1 be the nhumber of samples in D belonging to C_1 and N_2 be the number of samples in C_2 ($N_1 + N_2 = N$)

Likelihood function

$$P(\mathbf{t}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} [\pi \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{t_n} [(1-\pi)\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{(1-t_n)}$$

Sapienza University of Rome, Italy - Machine Learning (2018/2019)

Probabilistic Generative Models

Maximum likelihhod solution for 2 classes
Maximizing log likelihood function, we obtain

$$\pi = \frac{N_1}{N}$$

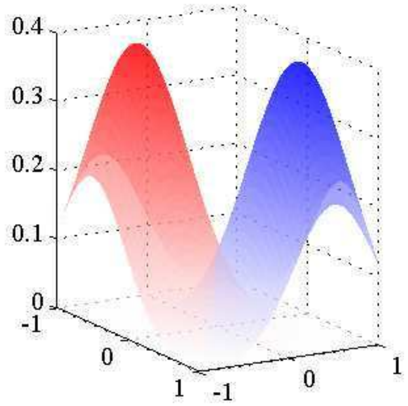
$$\boldsymbol{\mu}_1 = \frac{1}{N_1} \sum_{n=1}^{N} t_n \mathbf{x}_n \qquad \boldsymbol{\mu}_2 = \frac{1}{N_2} \sum_{n=1}^{N} (1 - t_n) \mathbf{x}_n$$

$$\boldsymbol{\Sigma} = \frac{N_1}{N} S_1 + \frac{N_2}{N} S_2$$

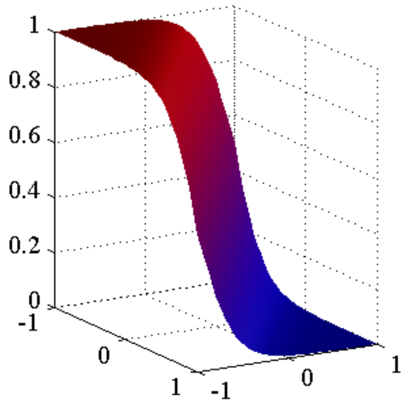
with $S_i = \frac{1}{N_i} \sum_{n \in C_i} (\mathbf{x}_n - \boldsymbol{\mu}_i)(\mathbf{x}_n - \boldsymbol{\mu}_i)^T$, $i = 1, 2$.

Note: details in C. Bishop. PRML. Section 4.2.2

Probabilistic Generative Models



$P(\mathbf{x}|C_1), P(\mathbf{x}|C_2)$



$P(C_1|\mathbf{x})$

Decision rule: $c = C_1 \iff P(c = C_1|\mathbf{x}) > 0.5$

Probabilistic Discriminative Models

Estimate directly $P(C_i|\mathbf{x})$

Logistic regression is a *classification* method based on maximum likelihood.

Logistic regression

Two classes

Given data set D, consider $\{\phi_n, t_n\}$, with $t_n \in \{0, 1\}$ and $\phi_n = \phi(\mathbf{x}_n)$, $n = 1, \ldots, N$

Likelihood function:

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n}(1 - y_n)^{1-t_n}$$

with $y_n = p(C_1|\phi_n) = \sigma(\mathbf{w}^T \phi_n)$

Logistic regression

Cross-entropy error function

$$E(\mathbf{w}) \equiv - \ln p(\mathbf{t}|\mathbf{w}) = - \sum_{n=1}^{N} [t_n \ln y_n + (1 - t_n) \ln(1 - y_n)]$$

Gradient of the error with respect to $\mathbf{w}$

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\phi_n$$

Iterative reweighted least squares

Apply *Newton-Raphson* iterative optimization for minimizing $E(\mathbf{w})$.

$$\mathbf{w} \leftarrow \mathbf{w} - \boldsymbol{H}^{-1} \nabla E(\mathbf{w})$$

$\boldsymbol{H} = \nabla\nabla E(\mathbf{w})$ is the Hessian matrix of $E(\mathbf{w})$ (second derivatives with respect to $\mathbf{w}$).

Sapienza University of Rome, Italy - Machine Learning (2018/2019)

Iterative reweighted least squares

$$\nabla E(\mathbf{w}) = \boldsymbol{\Phi}^T (\mathbf{y} - \mathbf{t})$$

$$\boldsymbol{H} = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1 - y_n)\phi_n \phi_n^T = \boldsymbol{\Phi}^T \boldsymbol{R} \boldsymbol{\Phi}$$

with $\mathbf{t} = (t_1, \ldots, t_n)^T$, $\mathbf{y} = (y_1, \ldots, y_n)^T$,
$\boldsymbol{R}$: diagonal matrix with $R_{nn} = y_n(1 - y_n)$,

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_1^T \\ \ldots \\ \phi_N^T \end{pmatrix}$$

Iterative reweighted least squares

Iterative method:

1. Initialize $\mathbf{w}$
2. Repeat until termination condition

$$\mathbf{w} \leftarrow \mathbf{w} - (\boldsymbol{\Phi}^T \boldsymbol{R} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T (\mathbf{y} - \mathbf{t})$$

Multiclass logistic regression

$$P(C_k|\phi) = y_k(\phi) = \frac{exp(a_k)}{\sum_j exp(a_j)}$$

with $a_k = \mathbf{w}_k^T \phi$.

Discriminative model

$$P(\mathbf{T}|\mathbf{w}_1, \ldots \mathbf{w}_K) = \prod_{n=1}^{N} \prod_{k=1}^{K} P(C_k|\phi_n)^{t_{nk}} = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

with $y_{nk} = y_k(\phi_n)$ and $\mathbf{T}$ $N \times K$ matrix of t_{nk}.

Multiclass logistic regression

Cross-entropy error function

$$E(\mathbf{w}_1, \dots \mathbf{w}_K) = -\ln P(\mathbf{T}|\mathbf{w}_1, \dots \mathbf{w}_K) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk}$$

Iterative algorithm with gradient $\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \dots \mathbf{w}_K) = ...$ and Hessian matrix $\nabla_{\mathbf{w}_k} \nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \dots \mathbf{w}_K) = ...$