Bayesian decision theory

Andrea Passerini
passerini@disi.unitn.it

Machine Learning

Introduction

Overview

- Bayesian decision theory allows to take optimal decisions in a fully probabilistic setting
- It assumes all relevant probabilities are known
- It allows to provide upper bounds on achievable errors and evaluate classifiers accordingly
- Bayesian reasoning can be generalized to cases when the probabilistic structure is not entirely known

Input-Output pair

Binary classification

- Assume examples $(x, y) \in \mathcal{X} \times \{-1, 1\}$ are drawn from a *known* distribution $p(x, y)$.
- The task is predicting the class y of examples given the input x.
- Bayes rule allows us to write it in probabilistic terms as:

$$P(y|x) = \frac{p(x|y)P(y)}{p(x)}$$

Output given input

Bayes rule

Bayes rule allows to compute the posterior probability given likelihood, prior and evidence:

$$posterior = \frac{likelihood \times prior}{evidence}$$

posterior
: $P(y|x)$ is the probability that class is y given that x was observed

likelihood
: $p(x|y)$ is the probability of observing x given that its class is y

prior
: $P(y)$ is the prior probability of the class, without any evidence

evidence
: $p(x)$ is the probability of the observation, and by the law of total probability can be computed as:

$$p(x) = \sum_{i=1}^{2} p(x|y)P(y)$$

Expected error

Probability of error

- Probability of error given x:

$$P(error|x) = \begin{cases} P(y_2|x) & \text{if we decide } y_1 \\ P(y_1|x) & \text{if we decide } y_2 \end{cases}$$

- Average probability of error:

$$P(error) = \int_{-\infty}^{\infty} P(error|x)p(x)dx$$

Bayes decision rule

Binary case

$$y_B = \text{argmax}_{y_i \in \{-1,1\}} P(y_i|x) = \text{argmax}_{y_i \in \{-1,1\}} p(x|y_i)P(y_i)$$

Multiclass case

$$y_B = \text{argmax}_{y_i \in \{1,\ldots,c\}} P(y_i|x) = \text{argmax}_{y_i \in \{1,\ldots,c\}} p(x|y_i)P(y_i)$$

Optimal rule

- The probability of error given x is:

$$P(error|x) = 1 - P(y_B|x)$$

- The Bayes decision rule minimizes the probability of error

Representing classifiers

Discriminant functions

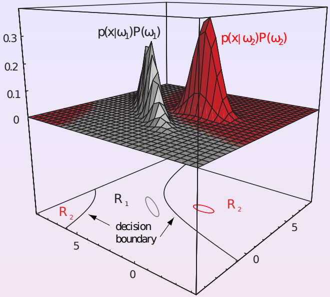
- A classifier can be represented as a set of *discriminant functions* $g_i(\mathbf{x}), i \in 1, \ldots, c$, giving:

$$y = \text{argmax}_{i \in 1,\ldots,c}\, g_i(\mathbf{x})$$

- A discriminant function is not unique $\Rightarrow$ the most convenient one for computational or explanatory reasons can be used:

$$g_i(\mathbf{x}) = P(y_i|\mathbf{x}) = \frac{p(\mathbf{x}|y_i)P(y_i)}{p(\mathbf{x})}$$
$$g_i(\mathbf{x}) = p(\mathbf{x}|y_i)P(y_i)$$
$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|y_i) + \ln P(y_i)$$

Representing classifiers



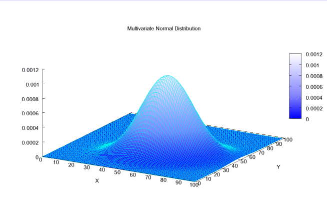
Decision regions

- The feature space is divided into *decision regions* $\mathcal{R}_1, \dots, \mathcal{R}_c$ such that:

$$\mathbf{x} \in \mathcal{R}_i \quad \text{if } g_i(\mathbf{x}) > g_j(\mathbf{x}) \; \forall j \neq i$$

- Decision regions are separated by *decision boundaries*, regions in which ties occur among the largest discriminant functions

Normal density

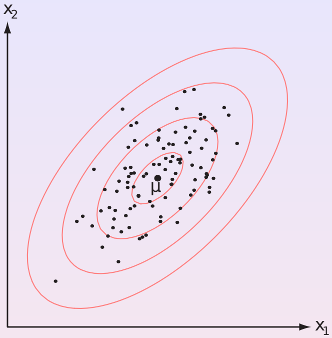


Multivariate normal density

$$\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

- The covariance matrix Σ is always symmetric and positive semi-definite
- The covariance matrix is strictly positive definite if the dimension of the feature space is d (otherwise $|\Sigma| = 0$)

Normal density



Hyperellipsoids

- The loci of points of constant density are hyperellipsoids of constant Mahalanobis distance from **x** to μ.
- The principal axes of such hyperellipsoids are the eigenvectors of Σ, their lengths are given by the corresponding eigenvalues

Discriminant functions for normal density

Discriminant functions

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|y_i) + \ln P(y_i)$$
$$= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \Sigma_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln |\Sigma_i| + \ln P(y_i)$$

Discarding terms which are independent of i we obtain:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \Sigma_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{1}{2}\ln |\Sigma_i| + \ln P(y_i)$$

Discriminant functions for normal density

case $\Sigma_i = \sigma^2 I$

- Features are statistically independent
- All features have same variance σ^2
- Covariance determinant $|\Sigma_i| = \sigma^{2d}$ can be ignored being independent of i
- Covariance inverse is given by $\Sigma_i^{-1} = (1/\sigma^2)I$
- The discriminant functions become:

$$g_i(\mathbf{x}) = -\frac{||\mathbf{x} - \boldsymbol{\mu}_i||^2}{2\sigma^2} + \ln P(y_i)$$

Discriminant functions for normal density

case $\Sigma_i = \sigma^2 I$

- Expansion of the quadratic form leads to:

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}[\mathbf{x}^t\mathbf{x} - 2\boldsymbol{\mu}_i^t\mathbf{x} + \boldsymbol{\mu}_i^t\boldsymbol{\mu}_i] + \ln P(y_i)$$

- Discarding terms which are independent of i we obtain *linear discriminant functions*:

$$g_i(\mathbf{x}) = \underbrace{\frac{1}{\sigma^2}\boldsymbol{\mu}_i^t}_{\boldsymbol{w}_i^t}\mathbf{x} \underbrace{-\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln P(y_i)}_{w_{i0}}$$

case $\Sigma_i = \sigma^2 I$

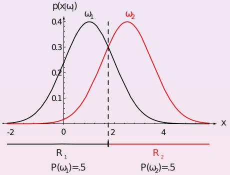
Separating hyperplane

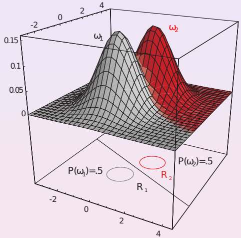
- Setting $g_i(\mathbf{x}) = g_j(\mathbf{x})$ we note that the decision boundaries are pieces of *hyperplanes*:

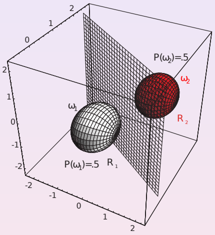
$$\underbrace{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t}_{\boldsymbol{w}^t}(\mathbf{x} - \left(\underbrace{\frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{||\boldsymbol{\mu}_i - \boldsymbol{\mu}_j||^2} \ln \frac{P(y_i)}{P(y_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}_{\mathbf{x}_0}\right))$$

- The hyperplane is orthogonal to vector $\boldsymbol{w} \Rightarrow$ orthogonal to the line linking the means
- The hyperplane passes through $\mathbf{x}_0$:
  - if the prior probabilities of classes are equal, $\mathbf{x}_0$ is halfway between the means
  - otherwise, $\mathbf{x}_0$ shifts away from the more likely mean

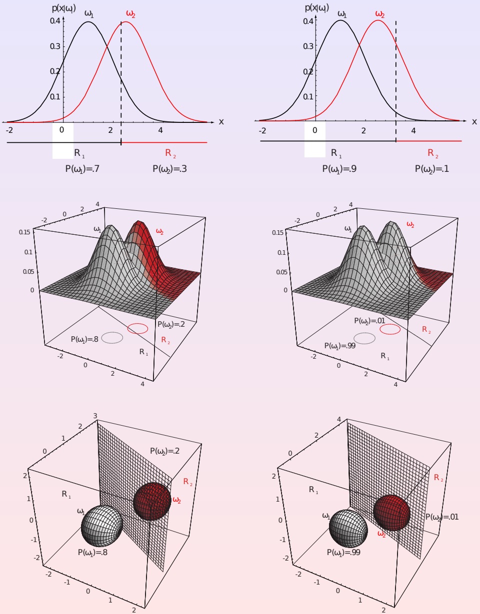
case $\Sigma_i = \sigma^2 I$



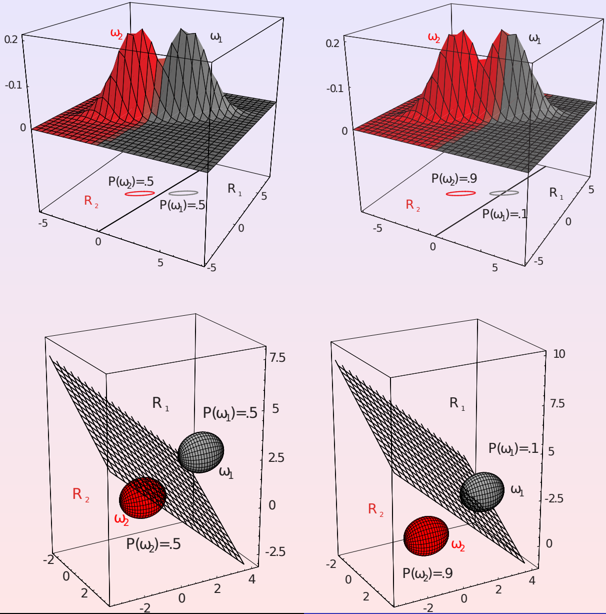




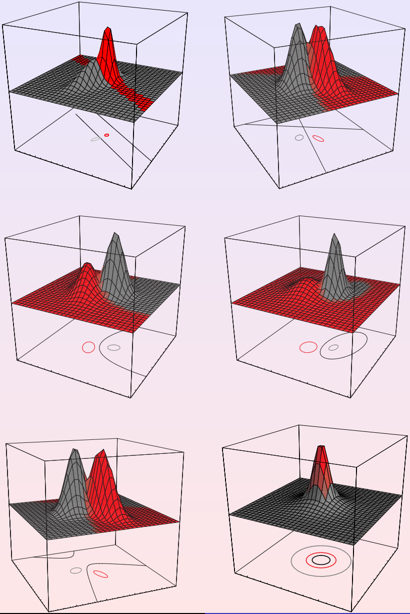
case $\Sigma_i = \sigma^2 I$



case $\Sigma_i = \Sigma$



case $\Sigma_i = arbitrary$



Appendix

Additional reference material

case $\Sigma_i = \sigma^2 I$

Separating hyperplane: derivation (1)

$$
\begin{aligned}
g_i(\mathbf{x}) - g_j(\mathbf{x}) &= 0 \\
\frac{1}{\sigma^2}\boldsymbol{\mu}_i^t\mathbf{x} - \frac{1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln P(y_i) - \frac{1}{\sigma^2}\boldsymbol{\mu}_j^t\mathbf{x} + \frac{1}{2\sigma^2}\boldsymbol{\mu}_j^t\boldsymbol{\mu}_j - \ln P(y_j) &= 0 \\
(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\mathbf{x} - 1/2(\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i - \boldsymbol{\mu}_j^t\boldsymbol{\mu}_j) + \sigma^2 \ln\frac{P(y_i)}{P(y_j)} &= 0 \\
\boldsymbol{w}^t(\mathbf{x} - \mathbf{x}_0) &= 0 \\
\boldsymbol{w} &= (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \\
(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\mathbf{x}_0 &= 1/2(\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i - \boldsymbol{\mu}_j^t\boldsymbol{\mu}_j) - \sigma^2 \ln\frac{P(y_i)}{P(y_j)}
\end{aligned}
$$

Separating hyperplane: derivation (2)

$$(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \mathbf{x}_0 = 1/2(\boldsymbol{\mu}_i^t \boldsymbol{\mu}_i - \boldsymbol{\mu}_j^t \boldsymbol{\mu}_j) - \sigma^2 \ln \frac{P(y_i)}{P(y_j)}$$

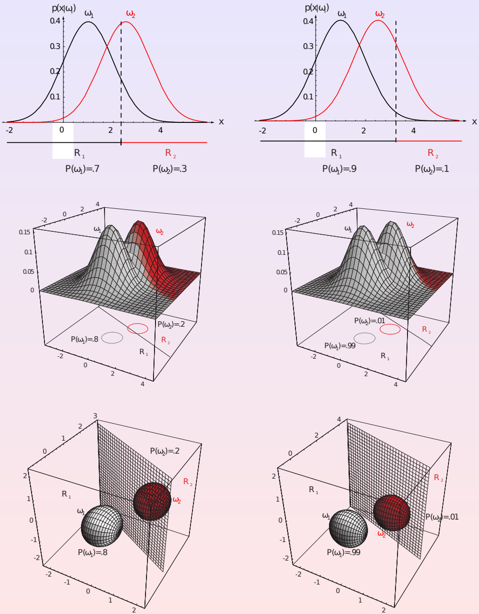
$$(\boldsymbol{\mu}_i^t \boldsymbol{\mu}_i - \boldsymbol{\mu}_j^t \boldsymbol{\mu}_j) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t (\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)$$

$$\ln \frac{P(y_i)}{P(y_j)} = \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)} \ln \frac{P(y_i)}{P(y_j)} =$$

$$= (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{||\boldsymbol{\mu}_i - \boldsymbol{\mu}_j||^2} \ln \frac{P(y_i)}{P(y_j)}$$

$$\mathbf{x}_0 = 1/2(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \sigma^2 \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{||\boldsymbol{\mu}_i - \boldsymbol{\mu}_j||^2} \ln \frac{P(y_i)}{P(y_j)}$$

case $\Sigma_i = \sigma^2 I$



Discriminant functions for normal density

case $\Sigma_i = \Sigma$

- All classes have same covariance matrix
- The discriminant functions become:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) + \ln P(y_i)$$

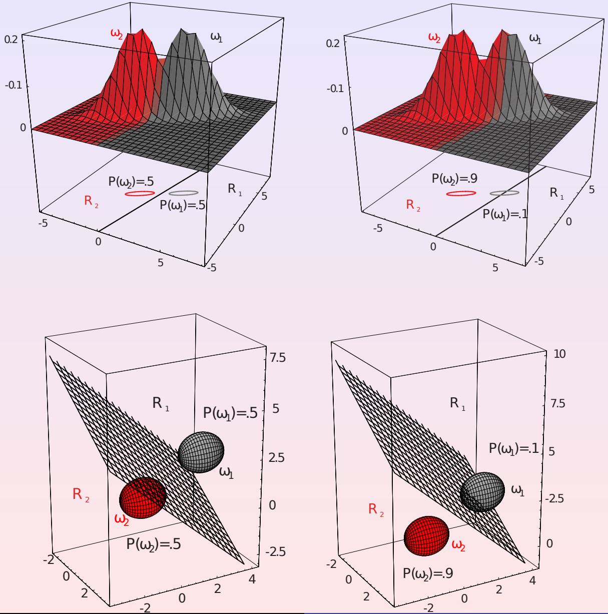
- Expanding the quadratic form and discarding terms independent of i we again obtain linear discriminant functions:

$$g_i(\mathbf{x}) = \underbrace{\boldsymbol{\mu}_i^t \Sigma^{-1}}_{\boldsymbol{w}_i^t} \mathbf{x} \underbrace{- \frac{1}{2}\boldsymbol{\mu}_i^t \Sigma^{-1} \boldsymbol{\mu}_i + \ln P(y_i)}_{w_{i0}}$$

- The separating hyperplanes are not necessarily orthogonal to the line linking the means:

$$\underbrace{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \Sigma^{-1}}_{\boldsymbol{w}^t} (\mathbf{x} - \underbrace{\frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln P(y_i)/P(y_j)}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \Sigma^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}_{\mathbf{x}_0})$$

case $\Sigma_i = \Sigma$



Discriminant functions for normal density

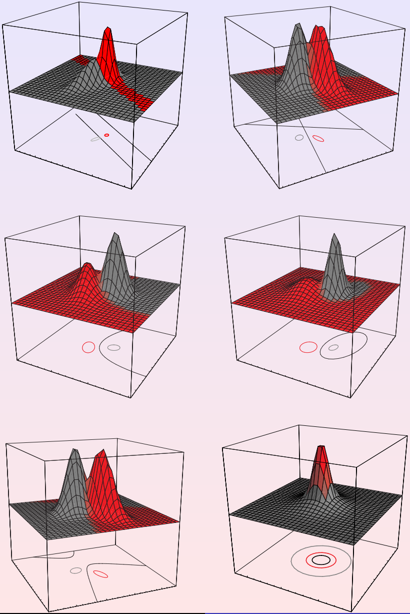
case Σ_i = *arbitrary*

- The discriminant functions are inherently quadratic:

$$g_i(\mathbf{x}) = \mathbf{x}^t \underbrace{(-\frac{1}{2}\Sigma_i^{-1})}_{W_i} \mathbf{x} + \underbrace{\boldsymbol{\mu}_i^t \Sigma_i^{-1}}_{\mathbf{w}_i^t} \mathbf{x} \underbrace{- \frac{1}{2}\boldsymbol{\mu}_i^t \Sigma_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(y_i)}_{w_{io}}$$

- In two category case, decision surfaces are *hyperquadratics*: hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, etc.

case Σ_i = *arbitrary*



Bayesian decision theory: arbitrary inputs and outputs

Setting

- Examples are input-output pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$ generated with probability $p(x, y)$.
- The *conditional risk* of predicting y^* given x is:

$$R(y^*|\mathbf{x}) = \int_{\mathcal{Y}} \ell(y^*, y) P(y|x) dy$$

- The overall *risk* of a decision rule f is given by

$$R[f] = \int R(f(x)|x) p(x) dx = \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell(f(x), y) p(y, x) dx dy$$

- Bayes decision rule

$$y^B = \mathrm{argmin}_{y \in \mathcal{Y}} R(y|x)$$

Handling missing features

Marginalize over missing variables

- Assume input $\mathbf{x}$ consists of an observed part $\mathbf{x}_o$ and missing part $\mathbf{x}_m$.
- Posterior probability of y_i given the observation can be obtained from probabilities over entire inputs by marginalizing over the missing part:

$$
\begin{aligned}
P(y_i|\mathbf{x}_o) &= \frac{p(y_i, \mathbf{x}_o)}{p(\mathbf{x}_o)} = \frac{\int p(y_i, \mathbf{x}_o, \mathbf{x}_m) d\mathbf{x}_m}{p(\mathbf{x}_o)} \\
&= \frac{\int P(y_i|\mathbf{x}_o, \mathbf{x}_m) p(\mathbf{x}_o, \mathbf{x}_m) d\mathbf{x}_m}{\int p(\mathbf{x}_o, \mathbf{x}_m) d\mathbf{x}_m} \\
&= \frac{\int P(y_i|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}_m}{\int p(\mathbf{x}) d\mathbf{x}_m}
\end{aligned}
$$

Handling noisy features

Marginalize over true variables

- Assume $\mathbf{x}$ consists of a clean part $\mathbf{x}_c$ and noisy part $\mathbf{x}_n$.
- Assume we have a *noise model* for the probability of the noisy feature given its true version $p(\mathbf{x}_n|\mathbf{x}_t)$.
- Posterior probability of y_i given the observation can be obtained from probabilities over clean inputs by marginalizing over true variables via the noise model:

$$\begin{aligned}
P(y_i|\mathbf{x}_c,\mathbf{x}_n) &= \frac{p(y_i,\mathbf{x}_c,\mathbf{x}_n)}{p(\mathbf{x}_c,\mathbf{x}_n)} = \frac{\int p(y_i,\mathbf{x}_c,\mathbf{x}_n,\mathbf{x}_t)d\mathbf{x}_t}{\int p(\mathbf{x}_c,\mathbf{x}_n,\mathbf{x}_t)d\mathbf{x}_t} \\
&= \frac{\int p(y_i|\mathbf{x}_c,\mathbf{x}_n,\mathbf{x}_t)p(\mathbf{x}_c,\mathbf{x}_n,\mathbf{x}_t)d\mathbf{x}_t}{\int p(\mathbf{x}_c,\mathbf{x}_n,\mathbf{x}_t)d\mathbf{x}_t} \\
&= \frac{\int p(y_i|\mathbf{x}_c,\mathbf{x}_t)p(\mathbf{x}_n|\mathbf{x}_c,\mathbf{x}_t)p(\mathbf{x}_c,\mathbf{x}_t)d\mathbf{x}_t}{\int p(\mathbf{x}_n|\mathbf{x}_c,\mathbf{x}_t)p(\mathbf{x}_c,\mathbf{x}_t)d\mathbf{x}_t} \\
&= \frac{\int p(y_i|\mathbf{x})p(\mathbf{x}_n|\mathbf{x}_t)p(\mathbf{x})d\mathbf{x}_t}{\int p(\mathbf{x}_n|\mathbf{x}_t)p(\mathbf{x})d\mathbf{x}_t}
\end{aligned}$$