Chapter 2: Bayesian Decision Theory

2.1 Bayes Theorem Definition

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

- X: **evidence** (observed sample)
- H: **hypothesis**
- $P(H)$: the **prior probability** that H holds
- $P(X|H)$: the **likelihood** of the evidence given the hypothesis
- $P(X)$: the **evidence probability**
- $P(H|X)$: the **posterior probability** that hypothesis H holds given evidence X

Bayes Theorem can be informally abstracted as

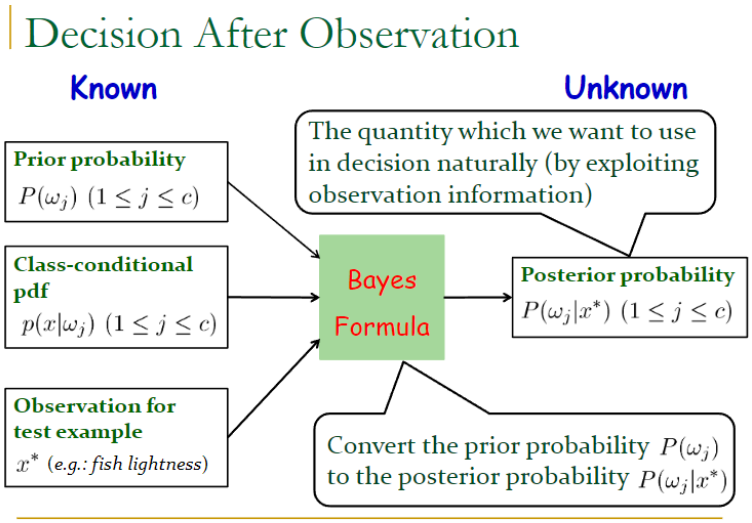
$$posterior = \frac{likelihood \times prior}{evidence}$$

2.2 Class-Conditional Probability Density Funciton

- a **probabilty density function (pdf)** for x given that the state of nature (class) is w, i.e.:

$$P(x|w) \geq 0, \quad \int_{-\infty}^{\infty} P(x|w)dx = 1$$

2.3 Bayes Formula



formula:

$$P(w_j|x) = \frac{P(x|w_j)P(w_j)}{\sum_{i=1}^{K} P(x|w_i)P(w_i)}$$

- for **constinous random variable**, given **joint probabiliyt density function** $p(w,x)$ and **marginal distribution** $P(w)$, $p(x)$

$$p(w) = \int_{-\infty}^{\infty} p(w,x)dx, \qquad p(x) = \sum_{j=1}^{K} p(w_j,x)$$

Bayes Decision Rule

$$if \quad P(w_j|x) \ge P(w_i|x), \forall j \ne i \quad \Rightarrow \quad w = w_j$$

- $P(x)$ is **irrelevant** for Bayesian decision

$$P(x) = \sum_{j=1}^{c} p(w_j,x) = \sum_{j=1}^{c} p(x|w_j)p(w_j)$$

2.4 Probalility of Error

In case of two classes, the **probability of error** is defined as

$$P(error|x) = \begin{cases} P(w_1|x), & \text{if we decide } w_2 \\ P(w_2|x), & \text{if we decide } w_1 \end{cases}$$

Under Bayes decision rule, we have

$$P(error|x) = \min[P(w_1|x), P(w_2|x)]$$

For every x, we ensure that $P(error|x)$ is as small as possible so that the **average probability of error** over all possible x must be as small as possible

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

2.5 The General Case

Bayes Decision Rule can be promoted to the following 4 formats:

1. Allowing to use **more than one feature**
2. Allowing to use **more than two states of nature (classes)**

3. Allowing **actions other than classification**
4. Introducing a **loss function** more general than the probability of error

loss function:

$$\lambda : \Omega \times \mathcal{A} \to \mathbb{R} \quad (\text{loss function})$$

- Ω: the set of all possible states of nature
- $\mathcal{A}$: the set of all possible actions
- $\lambda(w_j, \alpha_i)/\lambda(\alpha_i|w_j)$: the loss incurred for taking action α_i when the state of nature is w_j

Expected loss, also named ***(conditional) risk***:

- Average by **enumerating** over all possible states of nature:

$$\mathcal{R}(\alpha_i|x) = \mathbb{E}[\lambda(w, \alpha)] = \sum_{j=1}^{c} \lambda(\alpha_i|w_j)P(w_j|x)$$

  - $\lambda(\alpha_i|w_j)$: the loss incurred for taking action α_i when the state of nature is w_j
  - $P(w_j|x)$: the probability of the state of nature being w_j given the evidence x

Overall risk called ***Bayes risk*** (denoted as $\mathcal{R}$):

$$\mathcal{R} = \int \mathcal{R}(\alpha(x)|x)p(x)dx$$

- $\mathcal{R}(\alpha(x)|x)$: conditional risk for pattern x with action $\alpha(x)$
- $p(x)$: the probability density function for pattern x
- ***Bayes Decision Rule***(Genral case):

$$\alpha(x) = \arg\min_{\alpha \in \mathcal{A}} \mathcal{R}(\alpha(x)|x) = \arg\min_{\alpha \in \mathcal{A}} \sum_{j=1}^{c} \lambda(\alpha_i|w_j)P(w_j|x)$$

2.6 Two-Category Classification

Special Case:

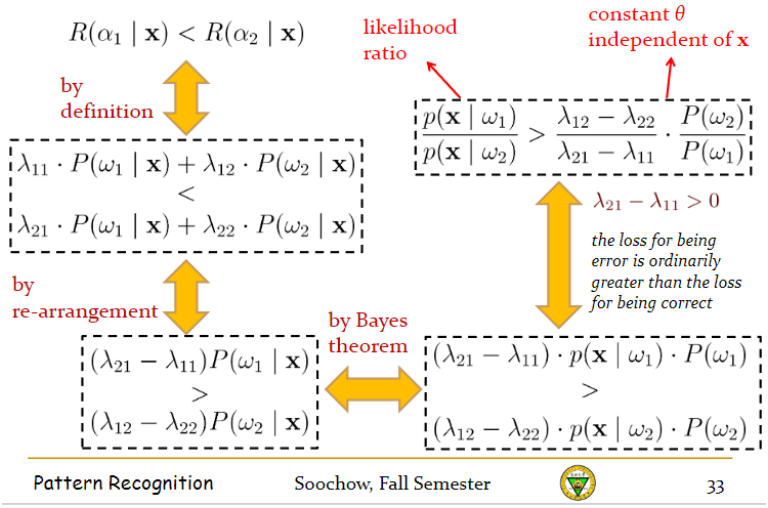
- $\Omega = \{w_1, w_2\}$ (two states of nature)
- $\mathcal{A} = \{\alpha_1, \alpha_2\}$ (α_i means deciding w_i)"
- $\lambda_{ij} = \lambda(\alpha_i|w_j)$: the loss incurred for deciding w_i when the true state of nature is w_j

The conditional risk:

$$\mathcal{R}(\alpha_1|x) = \lambda_{11}P(w_1|x) + \lambda_{12}P(w_2|x)$$
$$\mathcal{R}(\alpha_2|x) = \lambda_{21}P(w_1|x) + \lambda_{22}P(w_2|x)$$

- if $P(w_1|x) \geq P(w_2|x)$, decide w_1, else decide w_2

Let'u analyze the euqal rule:



- Generally **the loss for being error is ordinally larger than the loss for being correct**, so the factor $\lambda_{21} - \lambda_{11}$ and $\lambda_{12} - \lambda_{22}$ are both **positive**
- In the last inequality formula, formula $\frac{P(w_2)}{P(w_1)}$ is independent of x and can be seen as a constant, hence this mainly depends on the **probability density** of x. Given this format, Bayes Decision Rule can be explained as follows:
    - if the **likelihood ratio is larger than some threshold independent of the observed sample** x, decide w_1, else decide w_2

2.7 Minimum-Error-Rate Classification (Optional)

Classification setting

- $\Omega = \{w_1, w_2, ..., w_c\}$ (c possible states of nature)
- $\mathcal{A} = \{\alpha_1, \alpha_2 ..., \alpha_c\}$ (α_i means deciding w_i, $1 \leq i \leq c$)"

Zero-one (symmetrical) loss function

$$\lambda(\alpha_i|w_j) = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases} \qquad 1 \leq i, j \leq c$$

- Assign no loss to a correct decision
- Assign a unit loss to any incorrect decision (**equal cost**)

In this case, we can derive that:

$$\begin{aligned} \mathcal{R}(\alpha_i|x) &= \sum_{j=1}^{c} \lambda(\alpha_i|w_j) P(w_j|x) \\ &= \sum_{j \neq i} \lambda(\alpha_i|w_j) P(w_j|x) + \lambda(\alpha_i|w_i) P(w_i|x) \\ &= \sum_{j \neq i} P(w_j|x) \\ &= 1 - P(w_i|x) \end{aligned}$$

- $1 - P(w_i|x)$ (**error rate**): the probability that action α_i (decide w_i) is wrong

Minimum error rate:

$$\text{Decide } w_i \text{ if } P(w_i|x) \geq \max_{j=1}^{c} P(w_j|x)$$

2.8 Minimax Criterion

Varying prior probabilities leads to varying overall risk, and the **minimax criterion** aims to find the classifier which can **minimize the worst overall risk** for any value of the priors.

Two-category classification

- $\Omega = \{w_1, w_2\}$ (two states of nature)
- $\mathcal{A} = \{\alpha_1, \alpha_2\}$ (α_i means deciding w_i)"
- $\lambda_{ij} = \lambda(\alpha_i|w_j)$: the loss incurred for deciding w_i when the true state of nature is w_j

Suppose the two-category classfifier $\alpha(\cdot)$ decide w_1 in region $\mathcal{R}_1$ and decide w_2 in region $\mathcal{R}_2$. Here $\mathcal{R}_1 \cup \mathcal{R}_1 = \mathbf{R}^d$ and $\mathcal{R}_1 \cap \mathcal{R}_1 = \emptyset$. Then the **overall risk** is:

$$
\begin{aligned}
\mathcal{R} &= \int \mathcal{R}(\alpha(x)|x) \cdot p(x)dx \\
&= \int_{\mathcal{R}_1} \mathcal{R}(\alpha_1(x)|x) \cdot p(x)dx + \int_{\mathcal{R}_2} \mathcal{R}(\alpha_2(x)|x) \cdot p(x)dx \\
&= \int_{\mathcal{R}_1} \sum_{j=1}^{2} \mathcal{R}(\alpha_1|w_j) \cdot P(w_j|x) \cdot p(x)dx + \int_{\mathcal{R}_2} \sum_{j=1}^{2} \mathcal{R}(\alpha_2|w_j) \cdot P(w_j|x) \cdot p(x)dx \\
&= \int_{\mathcal{R}_1} \sum_{j=1}^{2} \mathcal{R}(\alpha_1|w_j) \cdot \frac{P(w_j) \cdot p(x|w_j)}{p(x)} \cdot p(x)dx + \int_{\mathcal{R}_2} \sum_{j=1}^{2} \mathcal{R}(\alpha_2|w_j) \cdot \frac{P(w_j) \cdot p(x|w_j)}{p(x)} \cdot p(x)dx \\
&= \int_{\mathcal{R}_1} \sum_{j=1}^{2} \lambda_{1j} \cdot P(w_j) \cdot p(x|w_j)dx + \int_{\mathcal{R}_2} \sum_{j=1}^{2} \lambda_{2j} \cdot P(w_j) \cdot p(x|w_j)dx \\
&= \int_{\mathcal{R}_1} \lambda_{11} \cdot P(w_1) \cdot p(x|w_1) + \lambda_{12} \cdot P(w_2) \cdot p(x|w_2)dx + \int_{\mathcal{R}_2} \lambda_{21} \cdot P(w_1) \cdot p(x|w_1) + \lambda_{22} \cdot P(w_2) \cdot p(x|w_2)dx
\end{aligned}
$$

Rewrite the overall risk $\mathcal{R}$ as a function of $P(w_1)$ via

- $P(w_2) = 1 - P(w_1)$
- $\int_{\mathcal{R}_1} p(x|w_1)dx = 1 - \int_{\mathcal{R}_2} p(x|w_1)dx$
- $\int_{\mathcal{R}_2} p(x|w_2)dx = 1 - \int_{\mathcal{R}_1} p(x|w_2)dx$

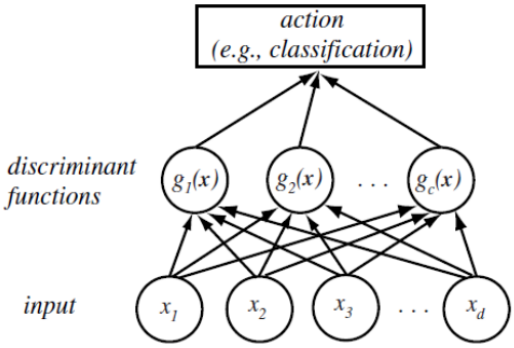
then we get:

$$
\begin{aligned}
\mathcal{R} &= \lambda_{11} \cdot P(w_1) \cdot [1 - \int_{\mathcal{R}_2} p(x|w_1)dx] + \int_{\mathcal{R}_1} \lambda_{12} \cdot [1 - P(w_1)] \cdot p(x|w_2)dx \\
&\quad + \overbrace{\int_{\mathcal{R}_2} \lambda_{21} \cdot P(w_1) \cdot p(x|w_1)dx + \lambda_{22} \cdot [1 - P(w_1)] \cdot [1 - \int_{\mathcal{R}_1} p(x|w_2)dx]}^{=\mathcal{R}_{mm}, \text{ minimax the overall risk}} \\
&= \lambda_{22} + (\lambda_{12} - \lambda_{22}) \cdot \int_{\mathcal{R}_1} p(x|w_2)dx \\
&\quad + P(w_1) \underbrace{\Big[(\lambda_{11} - \lambda_{22}) + (\lambda_{21} - \lambda_{11}) \cdot \int_{\mathcal{R}_2} p(x|w_1)dx - (\lambda_{12} - \lambda_{22}) \cdot \int_{\mathcal{R}_1} p(x|w_2)dx\Big]}_{=0, \text{ to solve the minimax situation}}
\end{aligned}
$$

we can see that the overall risk $\mathcal{R}$ is a linear function of $P(w_1)$. If we can find a decision boundary to ***make the ratio constant of*** $P(w_1)$ ***zero*** (i.e. $(\lambda_{11} - \lambda_{22}) + (\lambda_{21} - \lambda_{11}) \cdot \int_{\mathcal{R}_2} p(x|w_1)dx - (\lambda_{12} - \lambda_{22}) \cdot \int_{\mathcal{R}_1} p(x|w_2)dx = 0$), then we can **minimaxize risk** R_{mm}

- according to the aim **minimax the worst overall risk**, we look for the prior probability to make **bayes risks** biggest, the corresponding decision boundary give the result of **minimax decision**, hence **the risk value of minimax** $\mathcal{R}_{mm}$ **equals the worst Bayes risk**.

2.9 Discriminant Function



Discriminant functions: $g_i : \mathcal{R}^d \to \mathcal{R} \quad (1 \leq i \leq c)$

- Decide w_i, if $g_i(x) > g_j(x) \quad \forall j \neq i$
- Userful way to represent classifiers and one function per category

Usual risk under general situations, we use **Minimum risk** as the discriminant function: $g_i(x) = -\mathcal{R}(\alpha_i|x) \quad (i \leq i \leq c)$

- **Greatest discriminant function correspond with the smallest condition risk**

Under **minimum error probability** situations, we can simplify the problem by using $g_i(x) = P(w_i|x)$

- **Greatest discriminant function correspond with greatest posterior probability**

To simplify some problems, we introduce a **monotonically increasing function** $f(\cdot)$ and replace $g_i(x)$ with $f(g_i(x))$, which remain the classification result unchanged.

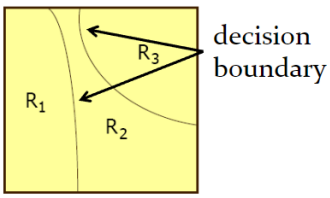
Decision region

$$\mathcal{R}_i = \{\mathbf{x} \in \mathbb{R}^d \mid g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \neq i\}$$

$$\text{where} \quad \mathcal{R}_i \cap \mathcal{R}_j = \emptyset \ (i \neq j) \quad \text{and} \quad \bigcup_{i=1}^{c} \mathcal{R}_i = \mathbb{R}^d$$

$$\mathbf{c} \text{ discriminant functions} \quad g_i(\cdot) \quad (1 \leq i \leq c) \quad \Longrightarrow \quad \mathbf{c} \text{ decision regions} \quad \mathcal{R}_i \subseteq \mathbb{R}^d \quad (1 \leq i \leq c)$$

Decision boundary: the surface in the decision space that maximizes the value of the discriminant function



2.10 Gaussian Density

Univariate Case, a.k.a. **normal density**, for continuous random variable:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{\frac{1}{2}\cdot(\frac{x-\mu}{\sigma})^2} \quad x \sim N(\mu, \sigma^2)$$

Vector Random Variable:

for $x = (x_1, x_2, ..., x_d)^T$

- **joint probability density function**: $x \sim p(x) = p(x_1, x_2, ..., x_d)$
- **marginal probability density function**: $p(x_1) = \int p(x_1, x_2) dx_2 \quad (x_1 \cap x_2 = \emptyset;\ x_1 \cup x_2 = x)$
- **expected vector**: $\varepsilon[x] = (\varepsilon[x_1], \varepsilon[x_2], ..., \varepsilon[x_d])^T$
    - $\varepsilon[x_i] = \int_{-\infty}^{\infty} x_i \cdot p(x_i) dx_i \quad (i \leq i \leq d)\backslash$

Covariance matrix

$$\Sigma = [\sigma_{ij}]_{1\le i,j\le d} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{pmatrix}$$

- **properties of Σ**
  - **symmetric**
  - **positive semi-definite**
- $$\begin{aligned} \sigma_{ij} = \sigma_{ji} &= \varepsilon\big[(x_i - \mu_i)(x_j - \mu_j)\big] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_i - \mu_i)(x_j - \mu_j)\cdot p(x_i, x_j)dx_i dx_j \end{aligned}$$
  - $\sigma_{ii} = Var[x_i] = \sigma_i^2$

Multivariate Case

$$x \sim N(\mu, \Sigma)$$
$$\mu_i = \varepsilon[x_i] \quad \sigma_{ij} = \sigma_{ji} = \varepsilon\big[(x_i - \mu_i)(x_j - \mu_j)\big]$$

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \cdot e^{-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)}$$

- d-dimensional column vector $x = (x_1, x_2, ..., x_d)^t$
- d-dimensional mean vector $\mu = (\mu_1, \mu_2, ..., \mu_d)^t$
- $d \times d$ covariance matrix

$$\Sigma = [\sigma_{ij}]_{1\le i,j\le d} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{pmatrix}$$

  - $|\Sigma|$: determinant of Σ
  - Σ^{-1}: inverse of Σ
- $\text{positive definite(artificial restricted) } \Sigma \Rightarrow \text{positive definite } \Sigma^{-1} \Rightarrow -\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu) < 0$

2.11 Discriminant Functions for Gaussian Density

According to **Minimum-error-rate classification**: $g_i(x) = P(w_i|x)$, we have discriminant function:

$$g_i(x) = P(w_i|x) \Longleftrightarrow g_i(x) = lnP(w_i|x) \Longleftrightarrow g_i(x) = lnp(x|w_i) + lnP(w_i|x)$$

If density function $p(x|w_i) \sim N(\mu_i, \Sigma)$ (**more esay for estimation**), we have discriminant function:

$$\begin{aligned} g_i(x) &= lnp(x|w_i) + lnP(w_i|x) \\ &= ln\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \cdot e^{-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)} + lnP(w_i) \\ &= -\frac{1}{2}(x-\mu_i)^t\Sigma^{-1}(x-\mu_i) - \overbrace{\frac{d}{2}\ln 2\pi}^{\text{constant}} - \frac{1}{2}ln|\Sigma_i| + lnP(w_i) \end{aligned}$$

Here are some special cases to discuss more specifically:

Case 1

$\sigma_i = \sigma^2 I$

Each feature is statistically independent with the same variance σ^2, hence we have

$$
\begin{aligned}
g_i(x) &= -\frac{1}{2}(x-\mu_i)^t\Sigma^{-1}(x-\mu_i) - \frac{1}{2}ln|\Sigma_i| + lnP(w_i) \\
&= -\frac{||x-\mu_i||}{2\sigma^2} + lnP(w_i) \quad (dln(\sigma) \text{ is constant, could be ignored}) \\
&= -\frac{1}{2\sigma^2}[\overbrace{x^tx}^{\text{same for all features}} - 2\mu_i^tx + \mu_i^t\mu_i] + lnP(w_i)
\end{aligned}
$$

The function above can be converted into an equivalent **linear discriminant function**:

$$
g_i(x) = w_i^tx + w_{i0}
$$

- **weight vector**: $w_i = \frac{1}{\sigma^2}\mu_i$
- **threshold/bias**: $w_{i0} = -\frac{1}{2\sigma^2}\mu_i^t\mu + lnP(w_i)$

Case 2

$\Sigma_i = \Sigma$

In this case, **the covariance matrix is the same for all classes**, hence we have:

$$
\begin{aligned}
g_i(x) &= -\frac{1}{2}(x-\mu_i)^t\Sigma^{-1}(x-\mu_i) + lnP(w_i) \\
&= -\frac{1}{2}(x^t-\mu_i^t)\Sigma^{-1}(x-\mu_i) + lnP(w_i) \\
&= -\frac{1}{2}(\overbrace{x^t\Sigma^{-1}x}^{\text{same for all features}} - 2\mu_i^t\Sigma^{-1}x + \mu_i^t\Sigma^{-1}\mu_i) + lnP(w_i)
\end{aligned}
$$

Converting it into a **linear discriminant function**:

$$
g_i(x) = w_i^tx + w_{i0}
$$

- **weight vector**: $w_i = \Sigma^{-1}\mu_i$
- **threshold/bias**: $w_{i0} = -\frac{1}{2}\mu_i^t\Sigma^{-1}\mu_i + lnP(w_i)$

Case 3

$\Sigma_i = \text{arbitrary}$

$$
\begin{aligned}
g_i(x) &= -\frac{1}{2}(x-\mu_i)^t\Sigma^{-1}(x-\mu_i) - \frac{1}{2}ln|\Sigma_i| + lnP(w_i) \\
&= -\frac{1}{2}(x^t-\mu_i^t)\Sigma^{-1}(x-\mu_i) - \frac{1}{2}ln|\Sigma_i| + lnP(w_i) \\
&= -\frac{1}{2}(x^t\Sigma^{-1}x - 2\mu_i^t\Sigma^{-1}x + \mu_i^t\Sigma^{-1}\mu_i) - \frac{1}{2}ln|\Sigma_i| + lnP(w_i)
\end{aligned}
$$

The function can be seen as a quadratic function, which is a **quadratic discriminant function**:

$$
g_i(x) = x^tW_ix + w_i^tx + w_{i0}
$$

- **quadratic matrix**: $W_i = -\frac{1}{2}\Sigma_i^{-1}$
- **weight vector**: $w_i = \Sigma_i^{-1}\mu_i$
- **threshold/bias**: $w_{i0} = -\frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i - \frac{1}{2}ln|\Sigma_i| + lnP(w_i)$