Bayesian Decision and Bayesian Learning

Ying Wu

Electrical Engineering and Computer Science
Northwestern University
Evanston, IL 60208

http://www.eecs.northwestern.edu/~yingwu

Bayes Rule

- $p(\mathbf{x}|\omega_i)$ *Likelihood*
- $p(\omega_i)$ *Prior*
- $p(\omega_i|\mathbf{x})$ *Posterior*
- Bayes Rule

$$p(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)p(\omega_i)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\omega_i)p(\omega_i)}{\sum_i p(\mathbf{x}|\omega_i)p(\omega_i)}$$

- In other words

$$\texttt{posterior} \propto \texttt{likelihood} \times \texttt{prior}$$

Outline

Action and Risk

- Classes: $\{\omega_1, \omega_2, \ldots, \omega_c\}$
- Actions: $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$
- Loss: $\lambda(\alpha_k|\omega_i)$
- Conditional risk:

$$R(\alpha_k|\mathbf{x}) = \sum_{i=1}^{c} \lambda(\alpha_k|\omega_i) p(\omega_i|\mathbf{x})$$

- Decision function, $\alpha(\mathbf{x})$, specifies a *decision rule*.
- Overall risk:

$$R = \int_x R(\alpha(\mathbf{x})|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

- It is the expected loss associated with a given decision rule.

Bayesian Decision and Bayesian Risk

- Bayesian decision

$$\alpha^* = \underset{k}{\mathit{argmin}}\, R(\alpha_k|\mathbf{x})$$

- This leads to the minimum overall risk. (why?)
- Bayesian risk: the minimum overall risk

$$R^* = \int_x R(\alpha^*|\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

- Bayesian risk is the best one can achieve.

Example: Minimum-error-rate classification

Let's have a specific example of Bayesian decision

- In classification problems, action α_k corresponds to ω_k
- Let's define a zero-one loss function

$$\lambda(\alpha_k|\omega_i) = \begin{cases} 0 & i = k \\ 1 & i \neq k \end{cases} \qquad i, k = 1, \ldots, c$$

  This means: no loss for correct decisions & all errors are equal

- It easy to see: the conditional risk $\rightarrow$ error rate

$$R(\alpha_k|\mathbf{x}) = \sum_{i \neq k} P(\omega_i|\mathbf{x}) = 1 - P(\omega_k|\mathbf{x})$$

- Bayesian decision rule $\rightarrow$ minimum-error-rate classification

$$\boxed{\textit{Decide } \omega_k \textit{ if } P(\omega_k|\mathbf{x}) > P(\omega_i|\mathbf{x}) \quad \forall i \neq k}$$

Outline

Classifier and Discriminant Functions

- Discriminant function: $g_i(\mathbf{x}), i = 1, \ldots, C$, assigns ω_i to $\mathbf{x}$
- Classifier

$$x \to \omega_i \ \textit{if} \ g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \neq i$$

- Examples:

$$\begin{aligned} g_i(\mathbf{x}) &= P(\omega_i|\mathbf{x}) \\ g_i(\mathbf{x}) &= P(\mathbf{x}|\omega_i)P(\omega_i) \\ g_i(\mathbf{x}) &= \ln P(\mathbf{x}|\omega_i) + \ln P(\omega_i) \end{aligned}$$

  Note: the choice of D-function is not unique, but they may give equivalent classification result.

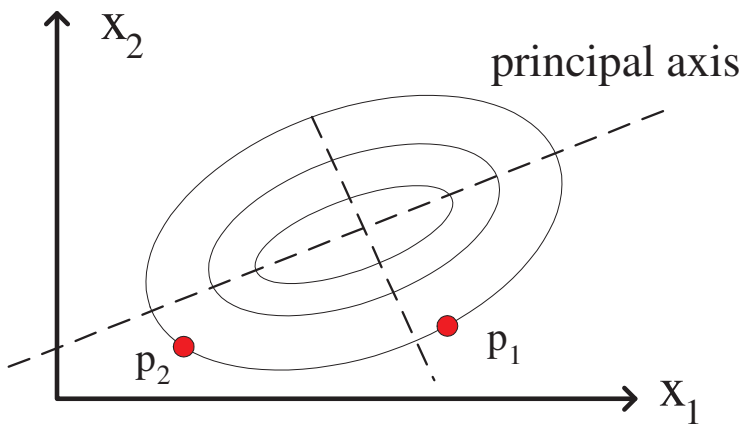
- Decision region: the partition of the feature space

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

- Decision boundary:

Multivariate Gaussian Distribution

$$p(\mathbf{x}) = N(\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

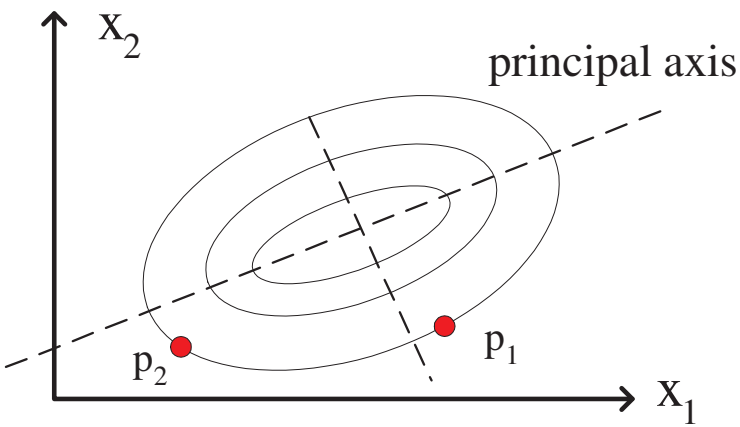


- The principal axes (the direction) are given by the eigenvectors of the covariance matrix Σ
- The length of the axes (the uncertainty) is given by the eigenvalues of Σ

Mahalanobis Distance

Mahalanobis distance is a normalized distance

$$||\mathbf{x} - \mu||_m = \sqrt{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}$$



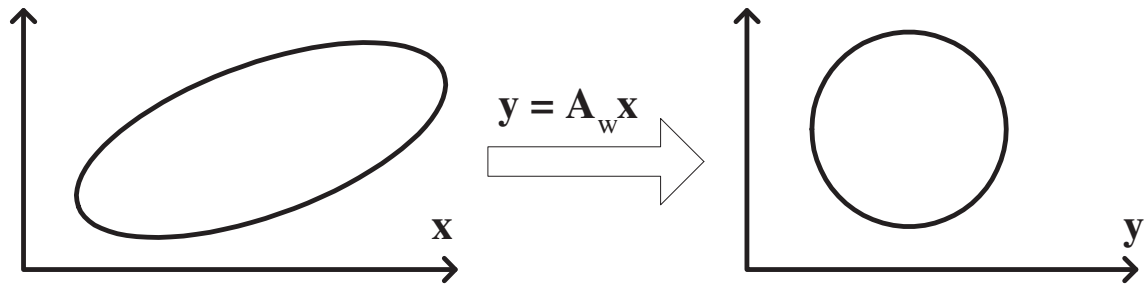
$$\begin{aligned} ||\mathbf{p}_1 - \mu||_2 &\neq ||\mathbf{p}_2 - \mu||_2 \\ ||\mathbf{p}_1 - \mu||_m &= ||\mathbf{p}_2 - \mu||_m \end{aligned}$$

Whitening

- To refresh your memory: A linear transformation of a Gaussian is still a Gaussian.

$$\begin{aligned} p(\mathbf{x}) &= N(\mu, \Sigma), \quad \textit{and } \mathbf{y} = \mathbf{A}^T\mathbf{x} \\ p(\mathbf{y}) &= N(\mathbf{A}^T\mu,\ \mathbf{A}^T\Sigma\mathbf{A}) \end{aligned}$$

- Question: Find one such that the covariance becomes an identity matrix (i.e., each component has equal uncertainty)



- Whitening is a transform that de-couples the correlation.

$$\mathbf{A}_w = \mathbf{U}^T\Lambda^{-\frac{1}{2}}, \quad \textit{where } \Sigma = \mathbf{U}^T\Lambda\mathbf{U}$$

- prove it: $\mathbf{A}_w^T\Sigma\mathbf{A}_w = \mathbf{I}$

Discriminant Functions for Gaussian Densities

Minimum-error-rate classifier

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln p(\omega_i)$$

When using Gaussian densities, it is easy to see:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma_i^{-1}(\mathbf{x} - \mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln p(\omega_i)$$

Case I: $\Sigma_i = \sigma^2 \mathbf{I}$

$$g_i(\mathbf{x}) = -\frac{||\mathbf{x} - \mu_i||^2}{2\sigma^2} + \ln p(\omega_i) = -\frac{1}{2\sigma^2}[\mathbf{x}^T\mathbf{x} - 2\mu_i^T\mathbf{x} + \mu_i^T\mu_i] + \ln p(\omega_i)$$

Notice that $\mathbf{x}^T\mathbf{x}$ is a constant. Equivalently we have

$$g_i(\mathbf{x}) = -[\frac{1}{\sigma^2}\mu_i]^T\mathbf{x} + [-\frac{1}{2\sigma^2}\mu_i^T\mu_i + \ln p(\omega_i)]$$

This leads to a *linear discriminant function*

$$g_i(\mathbf{x}) = \mathbf{W}_i^T\mathbf{x} + \mathbf{W}_{i0}$$

At the decision boundary $g_i(\mathbf{x}) = g_j(\mathbf{x})$, which is linear:

$$\mathbf{W}^T(\mathbf{x} - \mathbf{x}_0) = 0,$$

$$\textit{where} \quad \mathbf{W} = \mu_i - \mu_j \quad \textit{and}$$

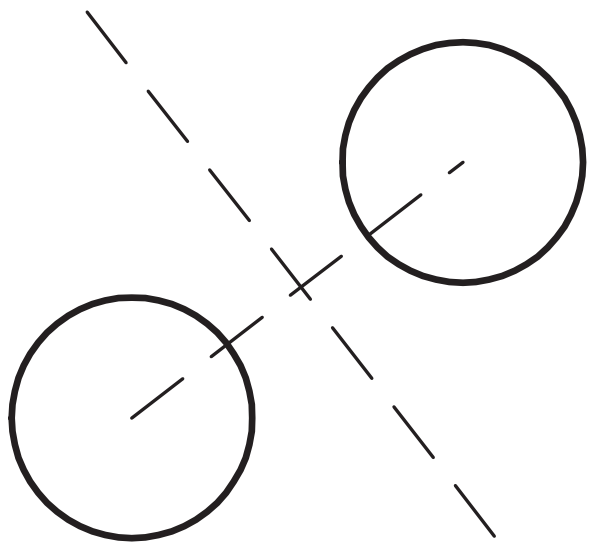
$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{||\mu_i - \mu_j||^2} \ln \frac{p(\omega_i)}{p(\omega_j)}(\mu_i - \mu_j)$$

To See it Clearly ...

Let's view a specific case, where $p(\omega_i) = p(\omega_j)$. The decision boundary we have:

$$(\mu_i - \mu_j)^T(\mathbf{x} - \frac{\mu_i + \mu_j}{2}) = 0$$

What does it mean?



The boundary is the perpendicular bisector of the two Gaussian densities!

what if $p(\omega_i) \neq p(\omega_j)$?

Case II: $\Sigma_i = \Sigma$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma^{-1}(\mathbf{x} - \mu_i) + \ln p(\omega_i)$$

Similarly, we can have an equivalent one:

$$g_i(\mathbf{x}) = (\Sigma^{-1}\mu_i)^T \mathbf{x} + (-\frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i + \ln(\omega_i))$$

The discriminant function and decision boundary are still linear:

$$\mathbf{W}^T(\mathbf{x} - \mathbf{x}_0) = 0$$

$$\textit{where} \quad \mathbf{W} = \Sigma^{-1}(\mu_i - \mu_j) \quad \textit{and}$$

$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln p(\omega_i) - \ln p(\omega_j)}{(\mu_i - \mu_j)^T \Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$$

Note: Compared with Case I, the Euclidean distance is replaced by Mahalanobis distance. The boundary is still linear, but the hyperplane is no longer orthogonal to $\mu_i - \mu_j$.

Case III: $\Sigma_i = \textit{arbitrary}$

$$g_i(\mathbf{x}) = \mathbf{x}^T \mathbf{A}_i \mathbf{x} + \mathbf{W}_i \mathbf{x} + \mathbf{W}_{i0},$$

where

$$\begin{aligned} \mathbf{A}_i &= -\frac{1}{2}\Sigma_i^{-1} \\ \mathbf{W}_i &= \Sigma_i^{-1}\mu_i \\ \mathbf{W}_{i0} &= -\frac{1}{2}\mu_i^T\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln p(\omega_i) \end{aligned}$$

Note: The decision boundary is no longer linear! It is byperquadrics.

Outline

Learning

- Learning means "training"
- i.e., estimating some unknowns from training samples
- Why?
  - It is very difficult to specify these unknowns
  - Hopefully, these unknowns can be recovered from examples given

Maximum Likelihood (ML) Estimation

- Collected samples $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$
- Estimate unknown parameters Θ in the sense that the data likelihood is maximized
- Data likelihood

$$p(\mathcal{D}|\Theta) = \prod_{k=1}^{n} p(x_k|\Theta)$$

- Log likelihood

$$L(\Theta) = \ln p(\mathcal{D}|\Theta) = \sum_{k=1}^{n} p(x_k|\Theta)$$

- ML estimation

$$\Theta^* = \underset{\Theta}{argmax}\, p(\mathcal{D}|\Theta) = \underset{\Theta}{argmax}\, L(\mathcal{D}|\Theta)$$

Example I: Gaussian densities (unknown μ)

$$\ln p(x_k|\mu) = -\frac{1}{2}\ln((2\pi)^d|\Sigma|) - \frac{1}{2}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)$$

Its partial derivative is:

$$\frac{\partial \ln p(x_k|\mu)}{\partial \mu} = \Sigma^{-1}(x_k-\mu)$$

So the KKT condition writes:

$$\sum_{k=1}^{n}\Sigma^{-1}(x_k-\hat{\mu}) = 0$$

It is easy to see the ML estimate of μ is:

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

This is exactly what we do in practice.

Example II: Gaussian densities (unknown μ and Σ)

Let's do the univariate case first. Denote σ^2 by Δ.

$$\ln p(x_k|\mu,\Delta) = -\frac{1}{2}\ln 2\pi\Delta - \frac{1}{2\Delta}(x_k-\mu)^2$$

The partial derivatives and KKT conditions are:

$$\begin{cases} \frac{\partial \ln p(x_k|\mu,\Delta)}{\partial \mu} = \frac{1}{\Delta}(x_k-\mu) \\ \frac{\partial \ln p(x_k|\mu,\Delta)}{\partial \Delta} = -\frac{1}{2\Delta} + \frac{(x_k-\mu)^2}{2\Delta^2} \end{cases} \Longrightarrow \quad \begin{cases} \sum_{k=1}^{n} \frac{1}{\hat{\Delta}}(x_k-\hat{\mu}) = 0 \\ \sum_{k=1}^{n}\{\frac{1}{\hat{\Delta}} + \frac{(x_k-\hat{\mu})^2}{\hat{\Delta}^2}\} = 0 \end{cases}$$

So we have

$$\begin{cases} \hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \\ \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k-\hat{\mu})^2 \end{cases} \overset{generalize}{\Longrightarrow} \begin{cases} \hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \\ \hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(x_k-\hat{\mu})(x_k-\hat{\mu})^T \end{cases}$$

Outline

Bayesian Estimation

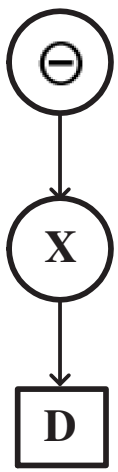
- Collect samples $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$, drawn independently from a fixed but unknown distribution $p(x)$
- Bayesian estimation uses $\mathcal{D}$ to determine $p(x|\mathcal{D})$, i.e., to learn a p.d.f.
- The unknown Θ is a random variable (or random vector), i.e., Θ is drawn from $p(\Theta)$.
- $p(\Theta)$ is unknown, but has a parametric form with parameters $\Theta \sim p(\Theta)$
- We hope $p(\Theta)$ is sharply peaked at the true value.
- Differences from ML
  - in Bayesian estimation, Θ is not a value, but a random vector
  - and we need to recover the distribution of Θ, rather than a single value.
  - $p(x|\mathcal{D})$ also needs to be estimated

Two Problems in Bayesian Learning

It is clear based on the total probability rule that

$$p(x|\mathcal{D}) = \int p(x, \Theta|\mathcal{D}) d\Theta = \int p(x|\Theta) p(\Theta|\mathcal{D}) d\Theta$$

- $p(x|\mathcal{D})$ is a weighted average over all Θ
- if $p(\Theta|\mathcal{D})$ peaks very sharply about some value $\hat{\Theta}$, then $p(x|\mathcal{D})$ can be approximated by $p(x|\hat{\Theta})$



- The generation of the observation $\mathcal{D}$ can be illustrated in a graphical model.
- The two problems are
  - estimating $p(\Theta|\mathcal{D})$
  - estimating $p(x|\mathcal{D})$

Example: The Univariate Gaussian Case $p(\mu|\mathcal{D})$

- Assume μ is the only unknown and it has a known Gaussian prior: $p(\mu) = N(\mu_0, \sigma_0^2)$.
- i.e., μ_0 is the best guess of μ, and σ_0 is its uncertainty
- Assume a Gaussian likelihood, $p(x|\mu) = N(\mu, \sigma^2)$
- It is clear that

$$p(\mu|\mathcal{D}) \sim p(\mathcal{D}|\mu)p(\mu) = \prod_{k=1}^{n} p(x_k|\mu)p(\mu)$$

$$\textit{where} \quad p(x_k|\mu) = N(\mu, \sigma^2) \quad \textit{and} \quad p(\mu) = N(\mu_0, \sigma_0^2)$$

- Let's prove that $p(\mu|\mathcal{D})$ is still a Gaussian density (why?)

$$\texttt{hint}: \quad p(\mu|\mathcal{D}) \sim \exp\{-\frac{1}{2}(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2})\mu^2 - 2(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma^2})\mu\}$$

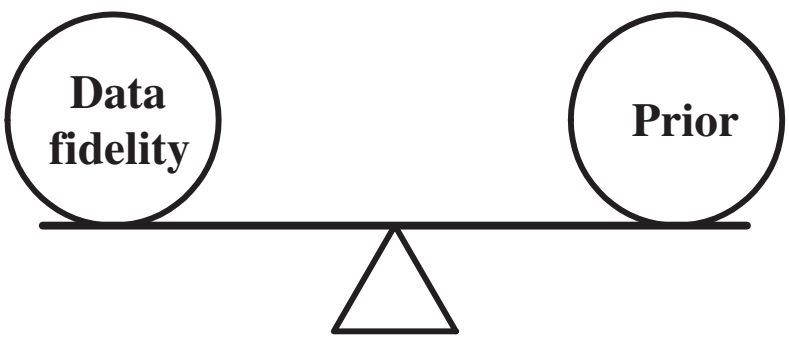
What is Going on?

As $\mathcal{D}$ is collection of n samples, let's denote $p(\mu|\mathcal{D}) = N(\mu_n, \sigma_n^2)$. Denote $\bar{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k$.

- μ_n represents our best guess for μ after observing n samples
- σ_n^2 measures the uncertainty of this guess
- So, what is really going on here?

We can obtain the following: (prove it!)

$$\begin{cases} \mu_n = \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\bar{\mu}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0 \\ \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2} \end{cases}$$



- $\bar{\mu}_n$ is the *data fidelity*
- μ_0 is the *prior*
- μ_n is a tradeoff (weighed average) between them

Example: The Univariate Gaussian case $p(x|\mathcal{D})$

After obtaining the posterior $p(\mu|\mathcal{D})$, we can estimate $p(x|\mathcal{D})$

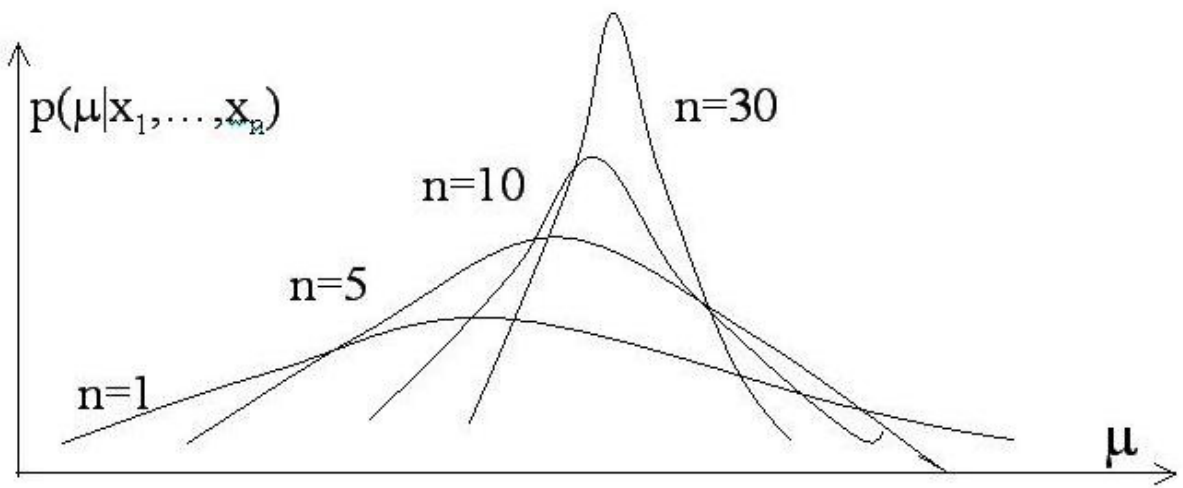
$$p(x|\mathcal{D}) = \int p(x|\mu)p(\mu|\mathcal{D})d\mu$$

It is the convolution of two Gaussian distributions. You can easily prove: do it!

$$p(x|\mathcal{D}) = N(\mu_n,\ \sigma^2 + \sigma_n^2)$$

What is really going on?

- Let's study σ_n
  - $\sigma_n^2 \downarrow$ monotonically
  - each additional observation decreases the uncertainty of the estimate
  - $\sigma_n^2 \rightarrow \frac{\sigma^2}{n} \rightarrow 0$
- Let's study $p(x|\mathcal{D})$
  - $p(x|\mathcal{D}) = N(\mu_n, \sigma^2 + \sigma_n^2)$
  - $p(\mu|\mathcal{D})$ becomes more and more sharply peaked



let's discuss

- if $\sigma_0 = 0$, then what?
- if $\sigma_0 \gg \sigma$, then what?

Example: The Multivariate case

We can generalize the univariate case to multivariate Gaussian, $p(x|\mu) = N(\mu, \Sigma),\ p(\mu) = N(\mu_0, \Sigma_0)$.

$$\begin{aligned} \mu_n &= \Sigma_0(\Sigma_0 + \frac{1}{n}\Sigma)^{-1}\bar{\mu}_n + \frac{1}{n}\Sigma(\Sigma_0 + \frac{1}{n}\Sigma)^{-1}\mu_0 \\ \Sigma_n &= \frac{1}{n}\Sigma_0(\Sigma_0 + \frac{1}{n}\Sigma)^{-1}\Sigma \end{aligned}$$

Actually, this is the best linear unbiased estimate (BLUE).

In addition,

$$p(x|\mathcal{D}) = N(\mu_n,\ \Sigma + \Sigma_n)$$

Recursive Learning

- Bayesian estimation can be done recursively, i.e., updating the previous estimates with new data.
- Denote $\mathcal{D}^n = \{x_1, x_2, \ldots, x_n\}$
- Easy to see

$$p(\mathcal{D}^n|\Theta) = p(x_n|\Theta)p(\mathcal{D}^{n-1}|\Theta), \quad \textit{and} \quad p(\Theta|\mathcal{D}^0) = p(\Theta)$$

- The recursion is:

$$p(\Theta|\mathcal{D}^n) = \frac{p(x_n|\Theta)p(\Theta|\mathcal{D}^{n-1})}{\int p(x_n|\Theta)p(\Theta|\mathcal{D}^{n-1})d\Theta}$$

- So we start from $p(\Theta)$, then move on to $p(\Theta|x_1)$, $p(\Theta|x_1, x_2)$, and so on