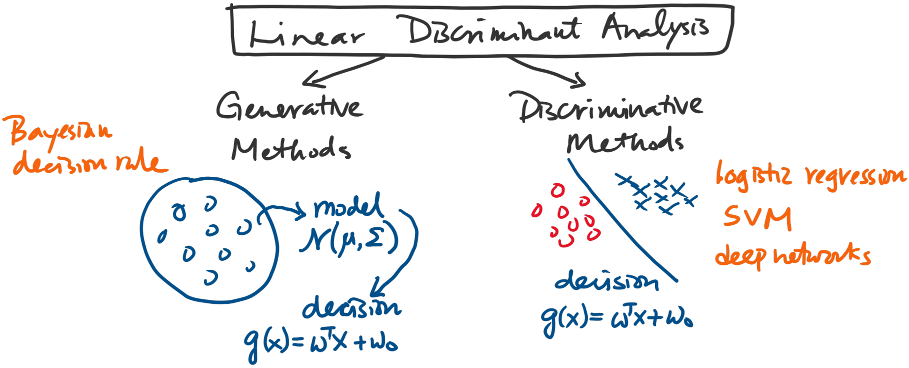
ECE595 / STAT598: Machine Learning I
Lecture 11 Maximum-Likelihood Estimation

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

Overview



- In linear discriminant analysis (LDA), there are generally two types of approaches
- **Generative approach**: Estimate model, then define the classifier
- **Discriminative approach**: Directly define the classifier

Outline

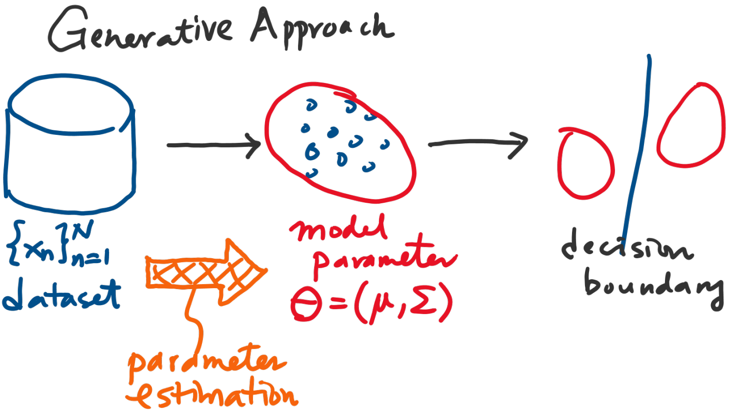
Generative Approaches

- Lecture 9 Bayesian Decision Rules
- Lecture 10 Evaluating Performance
- Lecture 11 Parameter Estimation
- Lecture 12 Bayesian Prior
- Lecture 13 Connecting Bayesian and Linear Regression

Today's Lecture

- Basic Principles
  - Likelihood Function
  - Maximum Likelihood Estimate
  - 1D Illustration
  - Gaussian Distributions
- Examples
  - Non-Gaussian Distributions
  - Biased and Unbiased Estimators
  - From MLE to MAP

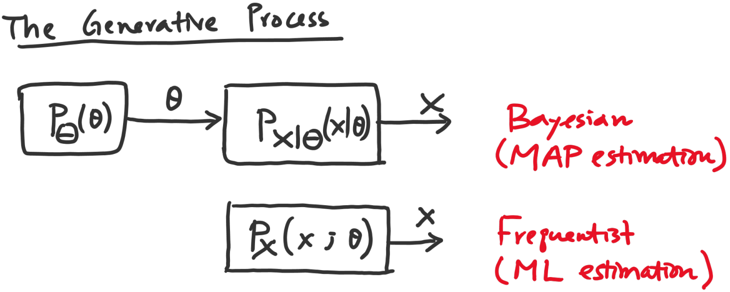
What is Parameter Estimation?



- The goal of parameter estimation is to determine $\boldsymbol{\Theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from dataset
- This is *the step* where you use data

MLE and MAP

There are two typical ways of estimating parameters.



- Maximum-likelihood estimation (MLE): $\boldsymbol{\theta}$ is deterministic.
- Maximum-a-posteriori estimation (MAP): $\boldsymbol{\theta}$ is random and has a prior distribution.

Maximum Likelihood Estimation

Given the dataset $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^N$, how to estimate the model parameters?

- We are going to use Gaussian as an illustration.
- Denote $\boldsymbol{\theta}$ as the model parameter.
- In Gaussian

$$\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$$

- The likelihood for one data point $\boldsymbol{x}_n$ is

$$p(\boldsymbol{x}_n \mid \overbrace{\boldsymbol{\theta}}^{=\{\boldsymbol{\mu},\boldsymbol{\Sigma}\}}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right\}$$

- $\boldsymbol{\theta}$ is a deterministic quantity, not a random variable.
- $\boldsymbol{\theta}$ does not have a distribution.
- $\boldsymbol{\theta}$ is fixed but unknown.

Likelihood for the Entire Dataset

- Likelihood for the entire dataset $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ is

$$
\begin{aligned}
p(\mathcal{D} \,|\, \boldsymbol{\theta}) &= \prod_{n=1}^{N} \left\{ \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right\} \right\} \\
&= \left(\frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \right)^{N} \exp\left\{ \sum_{n=1}^{N} -\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right\}
\end{aligned}
$$

- The Negative Log-Likelihood is

$$
\begin{aligned}
-\log p(\mathcal{D} \,|\, \boldsymbol{\theta}) &= \frac{N}{2} \log |\boldsymbol{\Sigma}| + \frac{N}{2} \log(2\pi)^d \\
&\quad + \sum_{n=1}^{N} \left\{ \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right\}.
\end{aligned}
$$

Maximum Likelihood Estimation

- Goal: Find $\boldsymbol{\theta}$ that maximizes the likelihood:

$$
\begin{aligned}
\widehat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\text{argmax}} \; p(\mathcal{D} \,|\, \boldsymbol{\theta}) \\
&= \underset{\boldsymbol{\theta}}{\text{argmax}} \; \prod_{n=1}^{N} \left\{ \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right\} \right\} \\
&= \underset{\boldsymbol{\theta}}{\text{argmin}} \; -\log(\cdots) \\
&= \underset{\boldsymbol{\theta}}{\text{argmin}} \; \frac{N}{2} \log|\boldsymbol{\Sigma}| + \frac{N}{2} \log(2\pi)^d \\
&\qquad\qquad + \sum_{n=1}^{N} \left\{ \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right\}.
\end{aligned}
$$

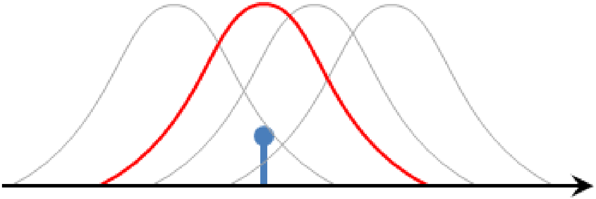
- This optimization is called the **maximum likelihood estimation** (MLE).

Illustrating MLE when $N = 1$. Known σ.

When $N = 1$: The MLE solution is

$$\begin{aligned}\widehat{\mu} &= \underset{\mu}{\text{argmax}} \; \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x_1-\mu)^2}{2\sigma^2} \right\} \\ &= \underset{\mu}{\text{argmin}} \; (x_1-\mu)^2 = x_1.\end{aligned}$$

- Which μ will give you the best Gaussian?
- When $\mu = x_1$, the probability of obtaining x_1 is the highest.



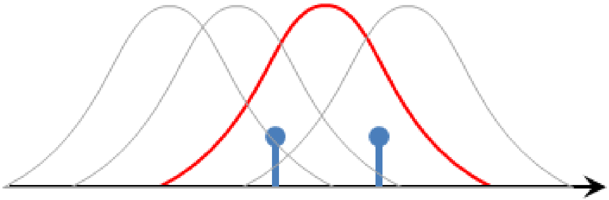
Illustrating MLE when $N = 2$. Known σ.

When $N = 2$: The MLE solution is

$$\widehat{\mu} = \underset{\mu}{\text{argmax}} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^2 \exp\left\{-\frac{(x_1-\mu)^2 + (x_2-\mu)^2}{2\sigma^2}\right\}$$

$$= \underset{\mu}{\text{argmin}} \;\; (x_1-\mu)^2 + (x_2-\mu)^2 = \frac{x_1+x_2}{2}.$$

- Which μ will give you the best Gaussian?
- When $\mu = (x_1 + x_2)/2$, the prob. of obtaining x_1 and x_2 is highest.



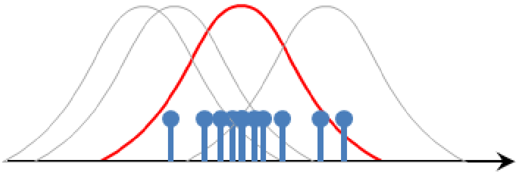
Illustrating MLE when N = arbitrary integer

The MLE solution is

$$\widehat{\mu} = \underset{\mu}{\operatorname{argmax}} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^2 \exp\left\{-\sum_{n=1}^{N} \frac{(x_n - \mu)^2}{2\sigma^2}\right\}$$

$$= \underset{\mu}{\operatorname{argmin}} \sum_{n=1}^{N} (x_n - \mu)^2 = \frac{1}{N}\sum_{n=1}^{N} x_n.$$

- Which μ will give you the best Gaussian?
- When $\mu = \frac{1}{N}\sum_{n=1}^{N} x_n$, the prob. of obtaining $\{x_n\}$ is highest.



Estimation in High-dimension

- Assume $\boldsymbol{\Sigma}$ is known and fixed.
- Thus, $\boldsymbol{\theta} = \boldsymbol{\mu}$. Estimate $\boldsymbol{\mu}$

$$
\begin{aligned}
\widehat{\boldsymbol{\mu}} = \underset{\boldsymbol{\mu}}{\operatorname{argmin}} \ & \cancel{\frac{N}{2} \log |\boldsymbol{\Sigma}|} + \cancel{\frac{N}{2} \log(2\pi)^d} \\
& \qquad + \sum_{n=1}^{N} \left\{ \frac{1}{2} (\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right\} \\
= \underset{\boldsymbol{\mu}}{\operatorname{argmin}} \ & \sum_{n=1}^{N} \left\{ (\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right\}
\end{aligned}
$$

- Take derivative, setting to zero:

$$
\nabla_{\boldsymbol{\mu}} \left\{ \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right\} = 2 \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) = \mathbf{0}.
$$

Estimation in High-dimension

- Let us do some algebra

$$\sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}) = \boldsymbol{0} \Longrightarrow \sum_{n=1}^{N} \boldsymbol{x}_n = \sum_{n=1}^{N} \boldsymbol{\mu}$$

- Then we can show that the MLE solution is

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n.$$

- This is just the **empirical average** of the entire dataset!
- You can show that if $\mathbb{E}[\boldsymbol{x}_n] = \boldsymbol{\mu}$ for all n, then

$$\mathbb{E}[\widehat{\boldsymbol{\mu}}] = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}[\boldsymbol{x}_n] = \boldsymbol{\mu}.$$

- We say that $\widehat{\boldsymbol{\mu}}$ is a **unbiased estimator** of $\boldsymbol{\mu}$ since $\mathbb{E}[\widehat{\boldsymbol{\mu}}] = \boldsymbol{\mu}$.

When both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are Unknown

What will be the MLE when both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown?

$$(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) = \underset{\boldsymbol{\mu}, \boldsymbol{\Sigma}}{\operatorname{argmin}} \; \underbrace{\frac{N}{2} \log |\boldsymbol{\Sigma}| + \frac{N}{2} \log(2\pi)^d + \sum_{n=1}^{N} \left\{ \frac{1}{2} (\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right\}}_{\varphi(\boldsymbol{\mu}, \boldsymbol{\Sigma})}.$$

You need to take derivative with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and solve

$$\begin{aligned} \nabla_{\boldsymbol{\mu}} \varphi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \boldsymbol{0} \\ \nabla_{\boldsymbol{\Sigma}} \varphi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \boldsymbol{0} \end{aligned}$$

With some (tedious) matrix calculus, we can show that

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n, \quad \text{and} \quad \widehat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^T.$$

Exercise: Prove this result when $\boldsymbol{x}_n$ is a 1D scalar.

Outline

Generative Approaches

- Lecture 9 Bayesian Decision Rules
- Lecture 10 Evaluating Performance
- Lecture 11 Parameter Estimation
- Lecture 12 Bayesian Prior
- Lecture 13 Connecting Bayesian and Linear Regression

Today's Lecture

- Basic Principles
  - Likelihood Function
  - Maximum Likelihood Estimate
  - 1D Illustration
  - Gaussian Distributions
- Examples
  - Non-Gaussian Distributions
  - Biased and Unbiased Estimators
  - From MLE to MAP

MLE does not need to be Gaussian

Suppose that $x_n \sim \text{Bernoulli}(\theta)$. Then,

$$p(x_n|\theta) = \begin{cases} \theta, & \text{if} \quad x_n = 1 \\ 1-\theta, & \text{if} \quad x_n = 0. \end{cases}$$

We can write the likelihood function as

$$\begin{aligned} p(\mathcal{D}|\theta) &= \prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{1-x_n} \\ &= \theta^{\sum_{n=1}^{N} x_n}(1-\theta)^{\sum_{n=1}^{N}(1-x_n)} \end{aligned}$$

and so the negative-log likelihood is

$$\begin{aligned} -\log p(\mathcal{D}|\theta) &= -\log\left\{\theta^{\sum_{n=1}^{N} x_n}(1-\theta)^{\sum_{n=1}^{N}(1-x_n)}\right\} \\ &= -\left(\sum_{n=1}^{N} x_n\right)\log\theta - \left(\sum_{n=1}^{N}(1-x_n)\right)\log(1-\theta) \end{aligned}$$

Bernoulli MLE

Taking the derivative and setting to zero:

$$\frac{d}{d\theta}\Big\{-\log p(\mathcal{D}|\theta)\Big\} = \frac{d}{d\theta}\left\{-\left(\sum_{n=1}^{N} x_n\right)\log\theta - \left(\sum_{n=1}^{N}(1-x_n)\right)\log(1-\theta)\right\}$$

$$= -\left(\sum_{n=1}^{N} x_n\right)\frac{1}{\theta} + \left(\sum_{n=1}^{N}(1-x_n)\right)\frac{1}{1-\theta}.$$

Setting to zero, we have

$$\begin{aligned}
\left(\textstyle\sum_{n=1}^{N} x_n\right)\tfrac{1}{\theta} &= \left(\textstyle\sum_{n=1}^{N}(1-x_n)\right)\tfrac{1}{1-\theta}\\
\left(\textstyle\sum_{n=1}^{N} x_n\right)(1-\theta) &= \left(\textstyle\sum_{n=1}^{N}(1-x_n)\right)\theta\\
\left(\textstyle\sum_{n=1}^{N} x_n\right) - \left(\textstyle\sum_{n=1}^{N} x_n\right)\theta &= N\theta - \left(\textstyle\sum_{n=1}^{N} x_n\right)\theta
\end{aligned}$$

Therefore,

$$\theta = \frac{1}{N}\sum_{n=1}^{N} x_n.$$

Unbias and Consistent Estimator

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n, \quad \text{and} \quad \widehat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_n - \widehat{\boldsymbol{\mu}})^T.$$

- $\widehat{\boldsymbol{\mu}}$ is the **empirical average** (or the sample mean)
- $\widehat{\boldsymbol{\Sigma}}$ is the **empirical covariance** (or the sample covariance)
- $\mathbb{E}[\widehat{\boldsymbol{\mu}}] = \boldsymbol{\mu}$
    - $\widehat{\boldsymbol{\mu}}$ is an **unbiased** estimate of $\boldsymbol{\mu}$: $\mathbb{E}[\widehat{\boldsymbol{\mu}}] = \boldsymbol{\mu}$ for all N
    - $\widehat{\boldsymbol{\mu}}$ is a consistent estimate of $\boldsymbol{\mu}$: As $N \to \infty$, $\widehat{\boldsymbol{\mu}} \xrightarrow{p} \boldsymbol{\mu}$
- $\mathbb{E}[\widehat{\boldsymbol{\Sigma}}] = \frac{N-1}{N}\boldsymbol{\Sigma}$
    - $\widehat{\boldsymbol{\Sigma}}$ is a **biased** estimate of $\boldsymbol{\Sigma}$: $\mathbb{E}[\widehat{\boldsymbol{\Sigma}}] \neq \boldsymbol{\Sigma}$
    - $\widehat{\boldsymbol{\Sigma}}$ is a consistent estimate of $\boldsymbol{\Sigma}$: As $N \to \infty$, $\widehat{\boldsymbol{\Sigma}} \xrightarrow{p} \boldsymbol{\Sigma}$
- You can make $\boldsymbol{\Sigma}$ unbiased by defining

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n, \quad \text{and} \quad \widehat{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{n=1}^{N}(\boldsymbol{x}_n - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_n - \widehat{\boldsymbol{\mu}})^T.$$

Unbiased and Consistent Estimator

Where does $(N-1)/N$ come from? Here is a 1D explanation. Assume $\mu = 0$.

$$\widehat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n, \quad \text{and} \quad \widehat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \widehat{\mu})^2$$

Taking expectation yields

$$\begin{aligned}
\mathbb{E}[\widehat{\sigma}^2] &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}(x_n - \widehat{\mu})^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x_n^2] - 2\mathbb{E}[\widehat{\mu}x_n] + \mathbb{E}[\widehat{\mu}^2] \\
&= \frac{1}{N}\sum_{n=1}^{N}\left\{\sigma^2 - 2\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}x_j x_n\right] + \mathbb{E}\left[\left(\frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right]\right\}
\end{aligned}$$

Unbiased and Consistent Estimator

$$\mathbb{E}[\widehat{\sigma}^2] = \frac{1}{N}\sum_{n=1}^{N}\left\{\sigma^2 - \underbrace{2\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} x_j x_n\right]}_{\substack{= \frac{2}{N}(\mathbb{E}[x_1 x_n + \ldots + x_N x_n]) \\ = \frac{2}{N}(0 + \ldots + \sigma^2 + \ldots + 0) \\ = \frac{2\sigma^2}{N}}} + \underbrace{\mathbb{E}\left[\left(\frac{1}{N}\sum_{n=1}^{N} x_n\right)^2\right]}_{\substack{= \frac{1}{N^2}\sum_{n=1}^{N}\mathbb{E}[x_n^2] + \sum_{j\neq n}\mathbb{E}[x_j x_n] \\ = \frac{1}{N^2} N\sigma^2 + 0 \\ = \frac{\sigma^2}{N}}}\right\}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left\{\sigma^2 - \frac{2}{N}\sigma^2 + \frac{1}{N}\sigma^2\right\}$$

$$= \frac{N-1}{N}\sigma^2.$$

From ML to Decision Boundary

- If you have a training set $\mathcal{D}$, ...
- Partition it according to the labels: $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$
- Pick a model, e.g., Gaussian
- For each class, estimate the model parameter
  - $\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1$ for $\mathcal{D}^{(1)}$
  - $\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2$ for $\mathcal{D}^{(2)}$
- Construct a discriminant function

$$g(\boldsymbol{x}) = \boldsymbol{x}^T(\boldsymbol{W}_1 - \boldsymbol{W}_2)\boldsymbol{x} + (\boldsymbol{w}_1 - \boldsymbol{w}_2)^T\boldsymbol{x} + (w_{10} - w_{20})$$

- See Lecture Bayesian Decision Rule 1 for formula
- Define the hypothesis function

$$h(\boldsymbol{x}) = \begin{cases} 1, & \text{if} \quad g(\boldsymbol{x}) > 0, \\ 0, & \text{if} \quad g(\boldsymbol{x}) < 0. \end{cases}$$

How well do you do?

- You have a dataset $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$.
- You estimate $\boldsymbol{\mu}$ by solving the maximum-likelihood problem.
- You get an estimate

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n$$

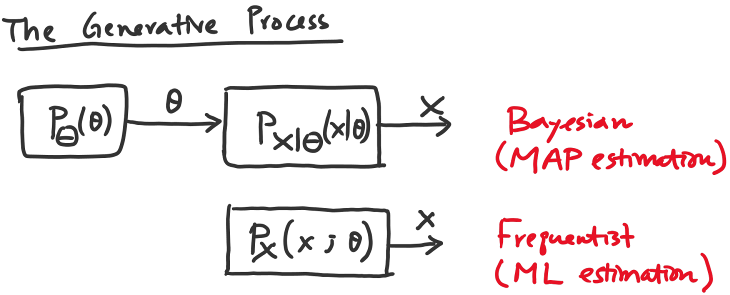
- How good is your estimate?
- Hoeffding inequality. In the 1D case:

$$\mathbb{P}\left[|\widehat{\mu} - \mu| > \epsilon\right] \leq 2e^{-2\epsilon^2 N}.$$

- As $N \to \infty$, $\widehat{\mu} \to \mu$ with very high probability.
- Overall performance:
  - Is $\widehat{\mu} \approx \mu$?
  - Is μ giving a good classifier?

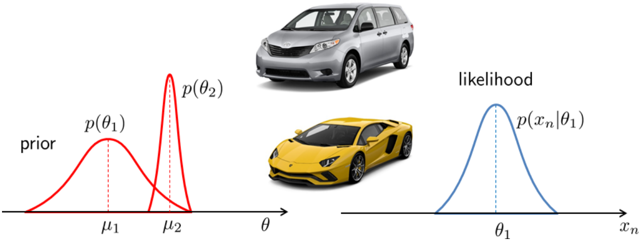
MLE and MAP

There are two typical ways of estimating parameters.



- Maximum-likelihood estimation (MLE): $\boldsymbol{\theta}$ is deterministic.
- Maximum-a-posteriori estimation (MAP): $\boldsymbol{\theta}$ is random and has a prior distribution.

From MLE to MAP



- Likelihood:

$$p(x_n|\theta_1) = \mathcal{N}(x_n|\theta_1, \sigma_1^2), \quad \text{and} \quad p(x_n|\theta_2) = \mathcal{N}(x_n|\theta_2, \sigma_2^2).$$

- Maximum-likelihood: You know nothing about θ_1 and θ_2. So you need to take measurements to estimate θ_1 and θ_2.
- Maximum-a-Posteriori: You know something about θ_1 and θ_2.
- Prior

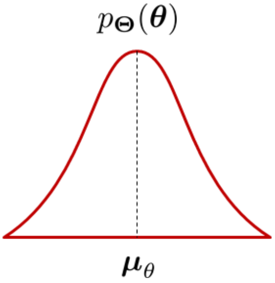
$$p(\theta_1) = \mathcal{N}(\mu_1|\gamma_1^2), \quad \text{and} \quad p(\theta_2) = \mathcal{N}(\mu_2|\gamma_2^2).$$

From MLE to MAP

- In MLE, the parameter $\boldsymbol{\theta}$ is **deterministic**.
- What if we assume $\boldsymbol{\theta}$ has a distribution?
- This makes $\boldsymbol{\theta}$ **probabilistic**.
- So make $\boldsymbol{\Theta}$ as a random variable, and $\boldsymbol{\theta}$ a state of $\boldsymbol{\Theta}$.
- Distribution of $\boldsymbol{\Theta}$:

$$p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$$

- $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ is the distribution of the parameter $\boldsymbol{\Theta}$.
- $\boldsymbol{\Theta}$ has its own mean and own variance.



Maximum-a-Posteriori

By Bayes Theorem again:

$$p_{\Theta|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}_n) = \frac{p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}_n|\boldsymbol{\theta})p_{\Theta}(\boldsymbol{\theta})}{p_{\boldsymbol{X}}(\boldsymbol{x}_n)}.$$

- To maximize the posterior distribution

$$\begin{aligned}
\widehat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \quad p_{\Theta|\boldsymbol{X}}(\boldsymbol{\theta}|\mathcal{D}) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \quad \prod_{n=1}^{N} p_{\Theta|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}_n) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \quad \prod_{n=1}^{N} \frac{p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}_n|\boldsymbol{\theta})p_{\Theta}(\boldsymbol{\theta})}{\cancel{p_{\boldsymbol{X}}(\boldsymbol{x}_n)}} \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \quad -\sum_{n=1}^{N}\left\{\log p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}_n|\boldsymbol{\theta}) + \log p_{\Theta}(\boldsymbol{\theta})\right\}
\end{aligned}$$

Reading List

Maximum Likelihood Estimation

- Duda-Hart-Stork, Pattern Classification, Chapter 3.2
- Iowa State EE 527 http://www.ece.iastate.edu/~namrata/EE527_Spring08/l5.pdf
- Purdue ECE 645, Lecture 18-20 https://engineering.purdue.edu/ChanGroup/ECE645.html
- UCSD ECE 271A, Lecture 6 http://www.svcl.ucsd.edu/courses/ece271A/ece271A.htm
- Univ. Orleans. https://www.univ-orleans.fr/deg/masters/ESA/CH/Chapter2_MLE.pdf