ECE595 / STAT598: Machine Learning I
Lecture 12.1: Bayesian Parameter Estimation - Basic Principles

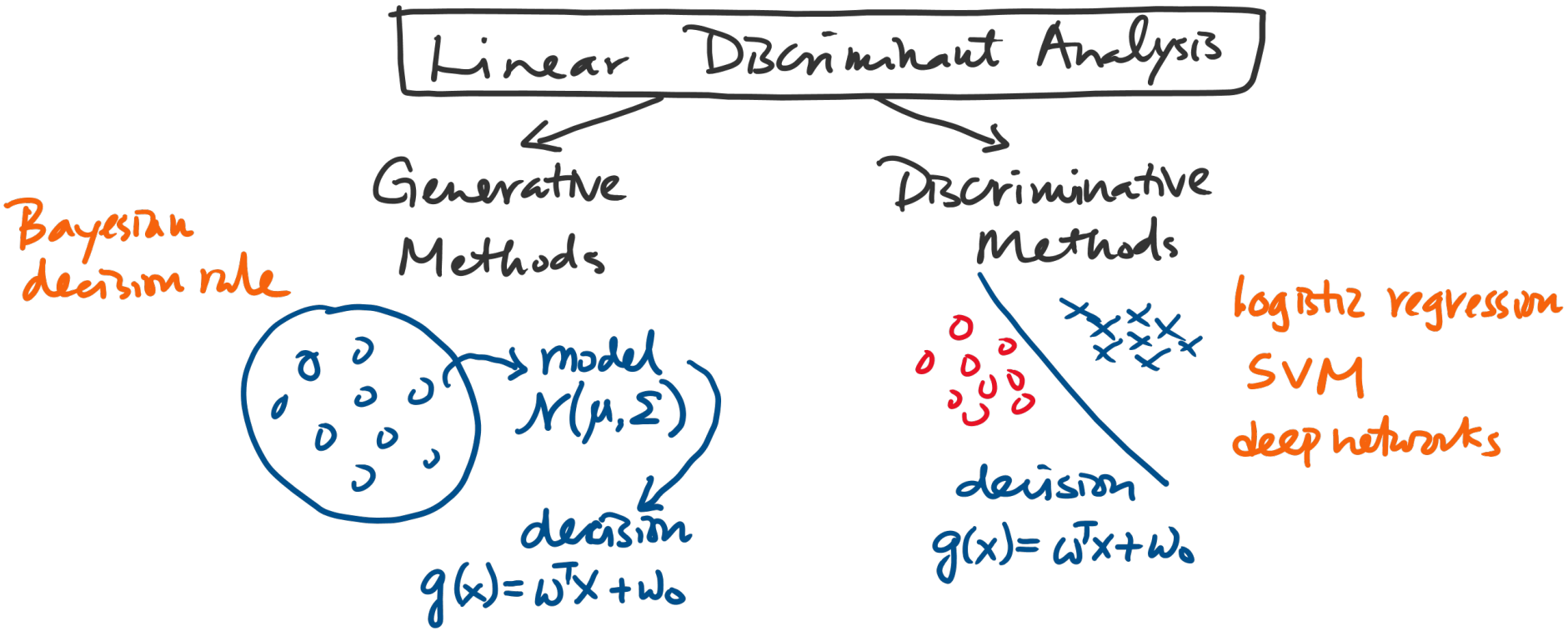
Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

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

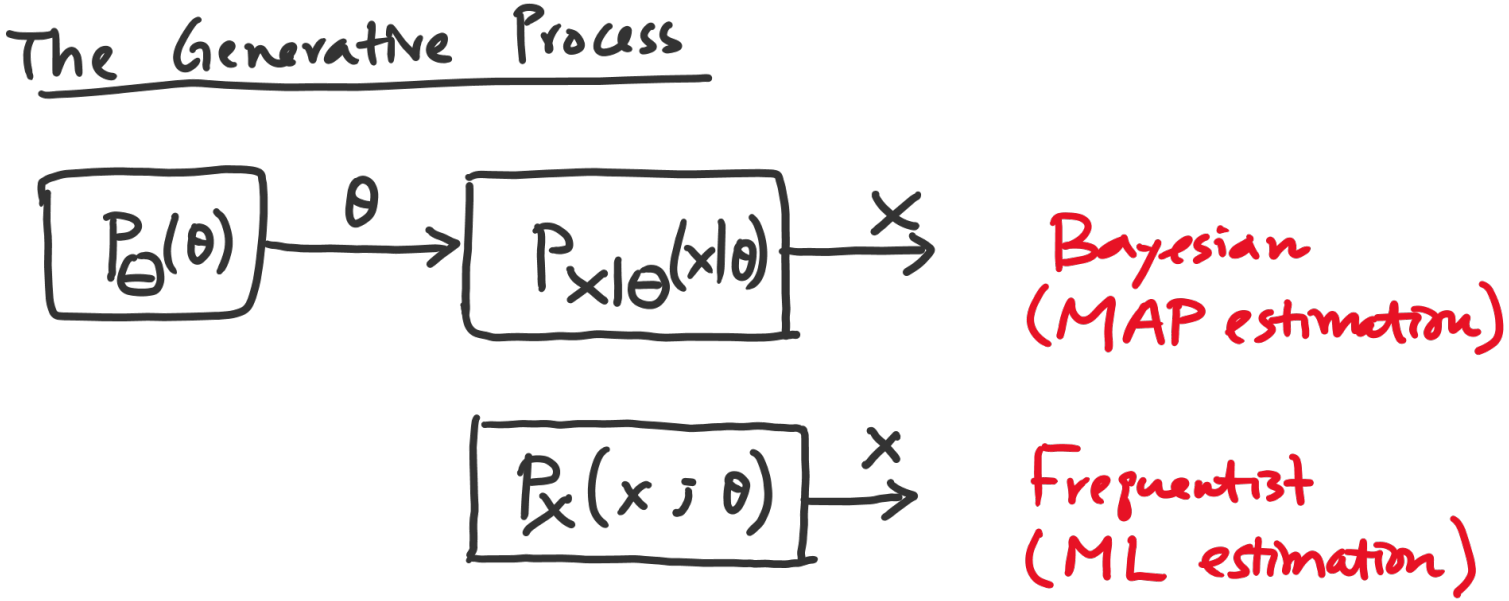
- Lecture 9 Bayesian Decision Rules
- Lecture 10 Evaluating Performance
- Lecture 11 Parameter Estimation
- Lecture 12 Bayesian Prior
- Lecture 13 Connecting Bayesian and Linear Regression

Today's Lecture

- Basic Principles
  - Posterior
  - 1D Illustration
  - Interpretations
- Choosing Priors
  - Prior for Mean
  - Prior for Variance
  - Conjugate Prior

MLE and MAP

There are two typical ways of estimating parameters.



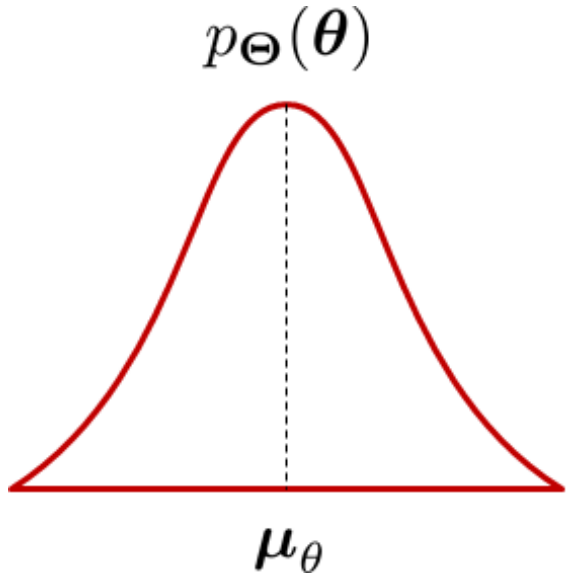
- Maximum-likelihood estimation (MLE): $\boldsymbol{\theta}$ is deterministic.
- Maximum-a-posteriori estimation (MAP): $\boldsymbol{\theta}$ is random and has a prior distribution.

From MLE to MAP

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

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

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



Maximum-a-Posteriori

By Bayes Theorem again:

$$p_{\Theta|X}(\theta|x_n) = \frac{p_{X|\Theta}(x_n|\theta)p_\Theta(\theta)}{p_X(x_n)}.$$

- To maximize the posterior distribution

$$\begin{aligned}
\widehat{\theta} &= \underset{\theta}{\text{argmax}} \quad p_{\Theta|X}(\theta|\mathcal{D}) \\
&= \underset{\theta}{\text{argmax}} \quad \prod_{n=1}^{N} p_{\Theta|X}(\theta|x_n) \\
&= \underset{\theta}{\text{argmax}} \quad \prod_{n=1}^{N} \frac{p_{X|\Theta}(x_n|\theta)p_\Theta(\theta)}{\cancel{p_X(x_n)}} \\
&= \underset{\theta}{\text{argmin}} \quad -\sum_{n=1}^{N} \Big\{ \log p_{X|\Theta}(x_n|\theta) + \log p_\Theta(\theta) \Big\}
\end{aligned}$$

The Role of $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$

- Let's look at the MAP:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \quad -\sum_{n=1}^{N} \Big\{ \log p_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x}_n|\boldsymbol{\theta}) + \log p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \Big\}$$

- Special case: When

$$p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0).$$

- Then the delta function gives

$$\log p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \begin{cases} -\infty, & \text{if} \quad \boldsymbol{\theta} \neq \boldsymbol{\theta}_0, \\ 0, & \text{if} \quad \boldsymbol{\theta} = \boldsymbol{\theta}_0. \end{cases}$$

- This will give

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \quad -\sum_{n=1}^{N} \begin{cases} -\infty, & \text{if} \quad \boldsymbol{\theta} \neq \boldsymbol{\theta}_0, \\ \log p_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x}_n|\boldsymbol{\theta}_0), & \text{if} \quad \boldsymbol{\theta} = \boldsymbol{\theta}_0. \end{cases} = \boldsymbol{\theta}_0.$$

- No uncertainty. Absolutely sure $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Illustration: 1D Example

Suppose that:

$$p_{X|\Theta}(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}$$

$$p_{\Theta}(\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right\}.$$

When $N = 1$. The MAP problem is simply

$$\begin{aligned}
\widehat{\theta} &= \underset{\theta}{\text{argmax}} \quad p_{X|\Theta}(x|\theta)p_{\Theta}(\theta) \\
&= \underset{\theta}{\text{argmax}} \quad \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\} \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right\} \\
&= \underset{\theta}{\text{argmax}} \quad -\frac{(x-\theta)^2}{2\sigma^2} - \frac{(\theta-\theta_0)^2}{2\sigma_0^2}
\end{aligned}$$

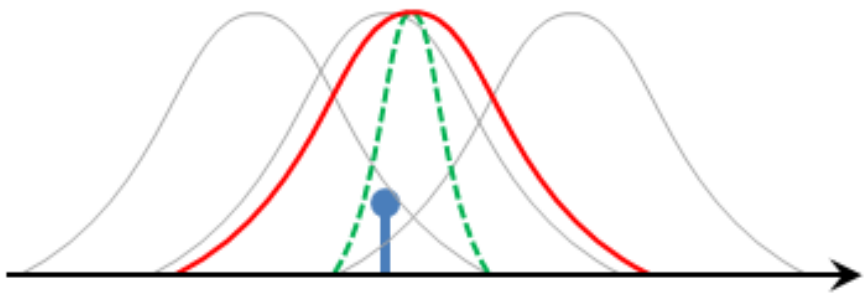
Illustration: 1D Example

Taking derivatives:

$$\begin{aligned}
& \frac{d}{d\theta}\left\{-\frac{(x-\theta)^2}{2\sigma^2}-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right\} && = 0 \\
\Rightarrow \quad & \frac{(x-\theta)}{\sigma^2}-\frac{(\theta-\theta_0)}{\sigma_0^2} && = 0 \\
\Rightarrow \quad & \sigma_0^2(x-\theta) && = \sigma^2(\theta-\theta_0) \\
\Rightarrow \quad & \sigma_0^2 x+\sigma^2\theta_0 && = (\sigma_0^2+\sigma^2)\theta
\end{aligned}$$

Therefore, the solution is

$$\theta = \frac{\sigma_0^2 x+\sigma^2\theta_0}{\sigma_0^2+\sigma^2}.$$



Interpreting the Result

Let us interpret the result

$$\theta = \frac{\sigma_0^2 x + \sigma^2 \theta_0}{\sigma_0^2 + \sigma^2}.$$

Does it make sense?

- If $\sigma_0 = 0$, then $\theta = \frac{\cancel{\sigma_0^2} x + \sigma^2 \theta_0}{\cancel{\sigma_0^2} + \sigma^2} = \theta_0$.
- This means: No uncertainty. Absolutely sure that $\theta = \theta_0$.
- $p_\Theta(\theta) = \delta(\theta - \theta_0)$

The other extreme

- If $\sigma_0 = \infty$, then $\theta = \frac{\sigma_0^2 x + \cancel{\sigma^2 \theta_0}}{\sigma_0^2 + \cancel{\sigma^2}} = x$.
- This means: I don't trust my prior at all. Use data.
- $p_\Theta(\theta) = \frac{1}{|\Omega|}$, for all $\theta \in \Omega$.

Therefore, MAP solution gives you a trade-off between data and prior.

When $N = 2$.

When $N = 2$. The MAP problem is

$$\begin{aligned}
\widehat{\theta} &= \operatorname*{argmax}_{\theta} \; p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)p_{\Theta}(\theta) \\
&= \operatorname*{argmax}_{\theta} \; \left(\prod_{n=1}^{2} p_{X|\Theta}(x_n|\theta)\right) p_{\Theta}(\theta) \\
&= \operatorname*{argmax}_{\theta} \; \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^2 \exp\left\{-\frac{(x_1-\theta)^2+(x_2-\theta)^2}{2\sigma^2}\right\} \\
&\qquad\qquad \times \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right\} \\
&= \operatorname*{argmin}_{\theta} \; -\log(\;\cdot\;) \\
&= \operatorname*{argmin}_{\theta} \; \left\{\frac{(x_1-\theta)^2}{2\sigma^2}+\frac{(x_2-\theta)^2}{2\sigma^2}+\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right\}
\end{aligned}$$

When $N = 2$.

Taking derivatives and setting to zero

$$\frac{d}{d\theta}\left\{\frac{(x_1-\theta)^2}{2\sigma^2}+\frac{(x_2-\theta)^2}{2\sigma^2}+\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right\}$$
$$=-\frac{x_1-\theta}{\sigma^2}-\frac{x_2-\theta}{\sigma^2}+\frac{\theta-\theta_0}{\sigma_0^2}=0.$$

Equating to zero yields

$$\theta=\frac{(x_1+x_2)\sigma_0^2+\theta_0\sigma^2}{2\sigma_0^2+\sigma^2}.$$

- If $\sigma_0 = 0$ (certain prior), then $\theta = \theta_0$.
- If $\sigma_0 = \infty$ (useless prior), then $\theta = \frac{x_1+x_2}{2}$.

When N is Arbitrary

General N.

$$\widehat{\theta} = \underset{\theta}{\text{argmax}} \left[\prod_{n=1}^{N} p_{X|\Theta}(x_n|\theta)\right] p_{\Theta}(\theta)$$

$$= \underset{\theta}{\text{argmax}} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left\{-\sum_{n=1}^{N}\frac{(x_n-\theta)^2}{2\sigma^2}\right\} \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right\}$$

$$= \underset{\theta}{\text{argmin}} \left\{\sum_{n=1}^{N}\frac{(x_n-\theta)^2}{2\sigma^2} + \frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right\}$$

$$= \frac{\sigma_0^2\sum_{n=1}^{N}x_n + \theta_0\sigma^2}{N\sigma_0^2+\sigma^2}.$$

What does it mean?

Interpreting the MAP solution: $N \to \infty$

Let's do some algebra:

$$\widehat{\theta} = \frac{(\sum_{n=1}^{N} x_n)\sigma_0^2 + \theta_0\sigma^2}{N\sigma_0^2 + \sigma^2} = \frac{(\sum_{n=1}^{N} x_n)\sigma_0^2 + \theta_0\sigma^2}{N(\sigma_0^2 + \frac{\sigma^2}{N})}$$

$$= \frac{\left(\frac{1}{N}\sum_{n=1}^{N} x_n\right)\sigma_0^2 + \frac{\sigma^2}{N}\theta_0}{\sigma_0^2 + \frac{\sigma^2}{N}}.$$

- Fix σ_0 and σ
- As $N \to \infty$,

$$\widehat{\theta} = \frac{\left(\frac{1}{N}\sum_{n=1}^{N} x_n\right)\sigma_0^2 + \cancel{\frac{\sigma^2}{N}\theta_0}}{\sigma_0^2 + \cancel{\frac{\sigma^2}{N}}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- This is the maximum-likelihood estimate.
- When I have a lot of samples, the prior does not really matter.

Interpreting the MAP solution: $N \to 0$

$$\widehat{\theta} = \frac{\left(\frac{1}{N}\sum_{n=1}^{N} x_n\right)\sigma_0^2 + \frac{\sigma^2}{N}\theta_0}{\sigma_0^2 + \frac{\sigma^2}{N}}$$

- Fix σ_0 and σ
- As $N \to 0$,

$$\widehat{\theta} = \frac{\cancel{\left(\frac{1}{N}\sum_{n=1}^{N} x_n\right)\sigma_0^2} + \frac{\sigma^2}{N}\theta_0}{\cancel{\sigma_0^2} + \frac{\sigma^2}{N}} = \theta_0.$$

- This is just the prior.
- When I have very few sample, I should rely on the prior.
- If the prior is good, then I can do well even if I have very few samples.
- Maximum-likelihood does not have the same luxury!

What Happens to the Posterior?

Consider

$$p(\mathcal{D}|\mu) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left\{-\sum_{n=1}^{N}\frac{(x_n-\mu)^2}{2\sigma^2}\right\}$$

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}.$$

We can show that the posterior is

$$p(\mu|\mathcal{D}) = \frac{1}{\sqrt{2\pi\sigma_N^2}}\exp\left\{-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right\},$$

where

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\text{ML}}$$

$$\sigma_N^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2+N\sigma_0^2}.$$

What Happens to the Posterior?

- x_n are generated from $\mu = 0.8$ and $\sigma^2 = 0.1$.
- When N increases, the posterior shifts towards the true distribution.