Deep Generative Models Background

Fall Semester 2024

René Vidal
Director of the Center for Innovation in Data Engineering and Science (IDEAS),
Rachleff University Professor, University of Pennsylvania
Amazon Scholar & Chief Scientist at NORCE

Outline

- **Basics of Probability, Statistics, Information Theory**
  - Discrete and Continuous Distributions, Independence
  - Marginals, Conditionals & Example for a Gaussian
  - Entropy, Mutual Information, KL Divergence
- Generative vs Discriminative Models
- Learning Generative Models
  - Learning Criterion: Maximum Likelihood Estimation
  - Learning Algorithm: Stochastic Gradient Descent
- Classes of Generative Models
  - Gaussian Models: Closed form Solution
  - General Models: Need for Structure
  - Taxonomy of Models
    - Latent variable models, Autoregressive models, Energy based models

Review of Probability and Statistics

- We define some basic notations
- Data $\boldsymbol{x} \in \mathbb{R}^D$ follows some data distribution $\boldsymbol{x} \sim p(\boldsymbol{x})$
- If $\boldsymbol{x}$ is discrete, then $p(\boldsymbol{x})$ is a probability mass function, taking on discrete values $k \in \mathcal{X} = \{1, \ldots, N\}$
- If $\boldsymbol{x}$ is continuous, then $p(\boldsymbol{x})$ is a probability density function
- Independence: $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent if and only if $p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y})$

Marginals, Conditionals

- Marginal distribution
  - In the continuous case

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y}$$

  - In the discrete case

$$p(\boldsymbol{x}) = \sum_{y} p(\boldsymbol{x}, \boldsymbol{y})$$

- Conditional distribution / Bayes rule

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

- Product rule

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{y}) &= p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y}) \\ &= p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) \end{aligned}$$

- Bayes rule

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

Marginal and Conditional Distribution for a Gaussian

- Assume $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \boldsymbol{\Sigma}_a & \boldsymbol{\Sigma}_c \\ \boldsymbol{\Sigma}_c^\top & \boldsymbol{\Sigma}_b \end{bmatrix}$$

- Then, we get the following results

$$p(\boldsymbol{x}_a) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a),$$

$$p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a|\widehat{\boldsymbol{\mu}}_a, \widehat{\boldsymbol{\Sigma}}_a), \text{ where}$$

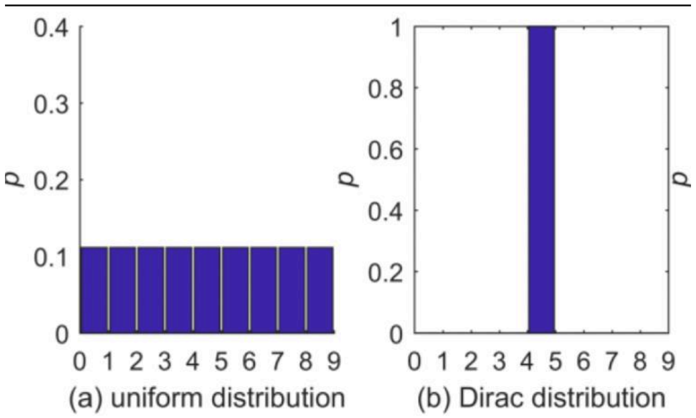
$$\widehat{\boldsymbol{\mu}}_a = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_c\boldsymbol{\Sigma}_b^{-1}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)),$$

$$\widehat{\boldsymbol{\Sigma}}_a = \boldsymbol{\Sigma}_a - \boldsymbol{\Sigma}_c\boldsymbol{\Sigma}_b^{-1}\boldsymbol{\Sigma}_c^\top$$

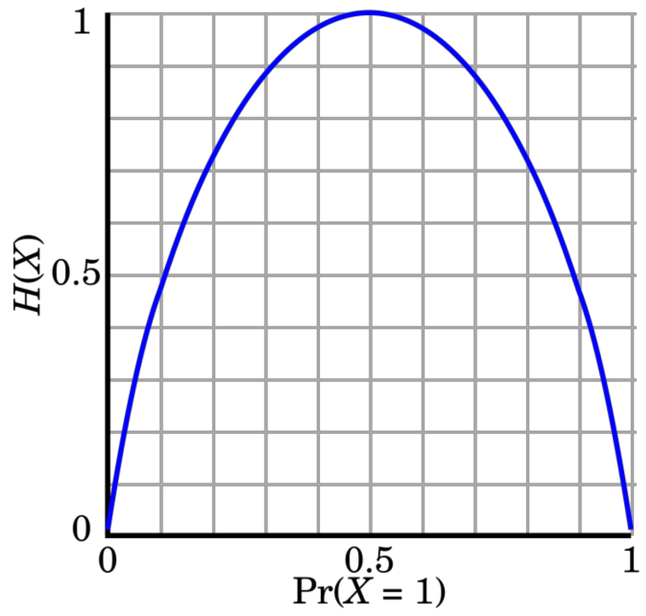
Warm-up exercise -> HW1

Review of Information Theory

- **Entropy** of a random variable X
  - It captures how much "uncertainty" is present in X
  - **Definition**: $H(X) = \mathrm{E}_{x \sim p(x)}[-\log p(x)]$
  - **Continuous**: $H(X) = -\int_x \log(p(x))\, p(x) dx$
  - **Discrete**: $H(X) = -\sum_{\mathrm{k}} \log(p_k) p_k$ where $p_k = P(X = k)$
  - **Example**: Let X be a Bernoulli random variable such that $\mathrm{P}(X = 1) = p$ and $\mathrm{P}(X = 0) = 1 - p$
    Then $H(X) = -p \log p - (1 - p) \log(1 - p)$



High entropy Low entropy



Entropy of a Bernoulli variable

- **Conditional entropy**: uncertainty of X when Y is observed
  - $H(X|Y) = \mathrm{E}_{x,y \sim p(x,y)}[-\log p(x|y)]$

Review of Information Theory

- **Mutual Information:**
  - mutual dependence between X and Y
  - reduction of uncertainty in X when Y is observed

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

  - $I(X;Y) = \mathrm{E}_{x,y\sim p(x,y)}\left[\log\left(\frac{p(x,y)}{p(x)p(y)}\right)\right]$
  - Note that if X, Y are independent, then $I(X;Y) = 0$
- **KL divergence** between two distributions p, q captures how similar p, q are

$$KL[p(x) \mid\mid q(x)] = E_{x\sim p}\left[\log\frac{p(x)}{q(x)}\right]$$

  - **Properties**
    - Non-negativity $KL[p(x) \mid\mid q(x)] \geq 0$. Equality holds iff $p = q$
    - In general triangle inequality and symmetry does not hold

Outline

- Basics of Probability, Statistics, Information Theory
  - Discrete and Continuous Distributions, Independence
  - Marginals, Conditionals & Example for a Gaussian
  - Entropy, Mutual Information, KL Divergence
- **Generative vs Discriminative Models**
- Learning Generative Models
  - Learning Criterion: Maximum Likelihood Estimation
  - Learning Algorithm: Stochastic Gradient Descent
- Classes of Generative Models
  - Gaussian Models: Closed form Solution
  - General Models: Need for Structure
  - Taxonomy of Models
    - Latent variable models, Autoregressive models, Energy based models

Statistical Generative Models

- A statistical generative model is a probability distribution $p(x)$

Image x  → A probability distribution $p(x)$ → probability $p(x)$

- It is generative because **sampling from $p(x)$ generates new images**

 … 

Discriminative vs. Generative Models

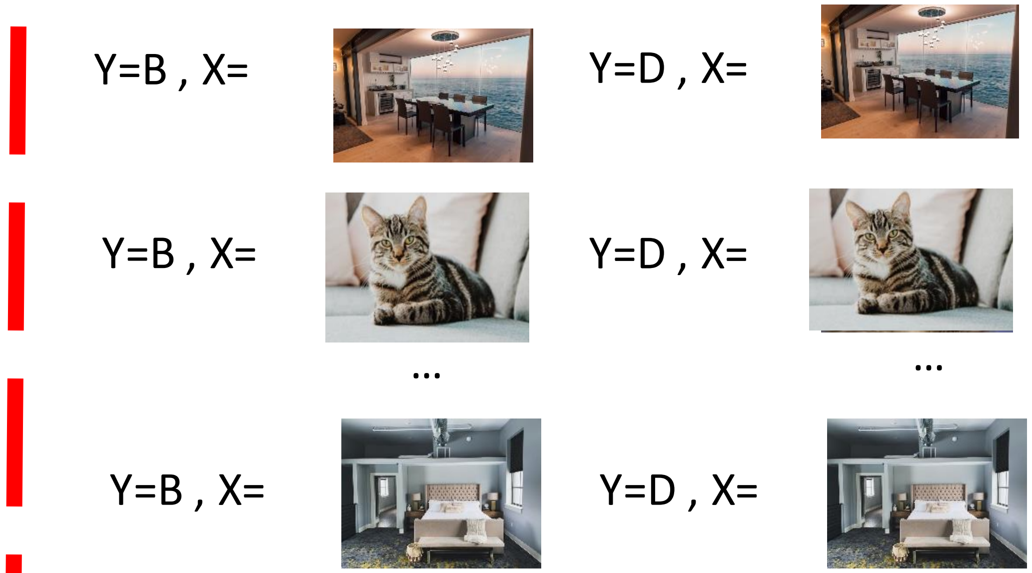
Discriminative: classify bedroom vs. dining room

The image X is given. **Goal**: decision boundary, via **conditional distribution over label Y**

P(Y = Bedroom | X= ) = 0.0001

Ex: logistic regression, convolutional net, etc.

Generative: generate X



The input X is **not** given. Requires a model of the **joint distribution over both X and Y**

P(Y = Bedroom , X= ) = 0.0002

Discriminative vs. Generative

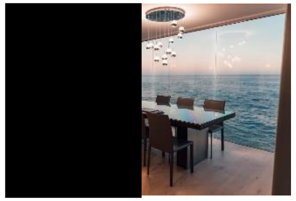
Joint and conditional are related via **Bayes Rule**:

P(Y = Bedroom | X= ) =

P(Y = Bedroom, X= )

P(X = )

Discriminative: Y is simple; X is always given, so not need to model P(X= )

Therefore it cannot handle missing data P(Y = Bedroom | X = )

Conditional Generative Models

Class **conditional generative models** are also possible:

P(X=  | Y = Bedroom)

It's often useful to condition on rich side information Y

P(X=  | Caption = "A black table with 6 chairs")

A discriminative model is a very simple conditional generative model of Y:

P(Y = Bedroom | X= )

Why Generative Models?

- AI Is Not Only About Decision Making

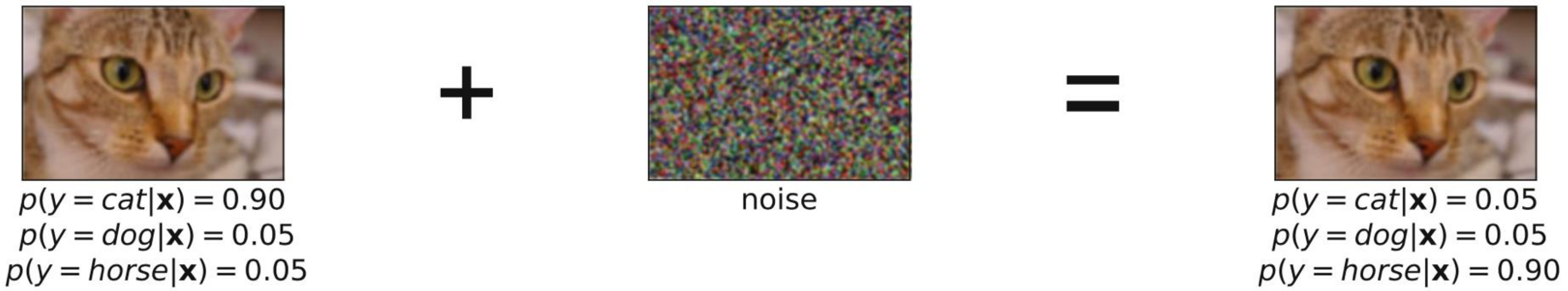


Fig. 1.1 An example of adding noise to an almost perfectly classified image that results in a shift of predicted label

- Importance of uncertainty and understanding in decision making

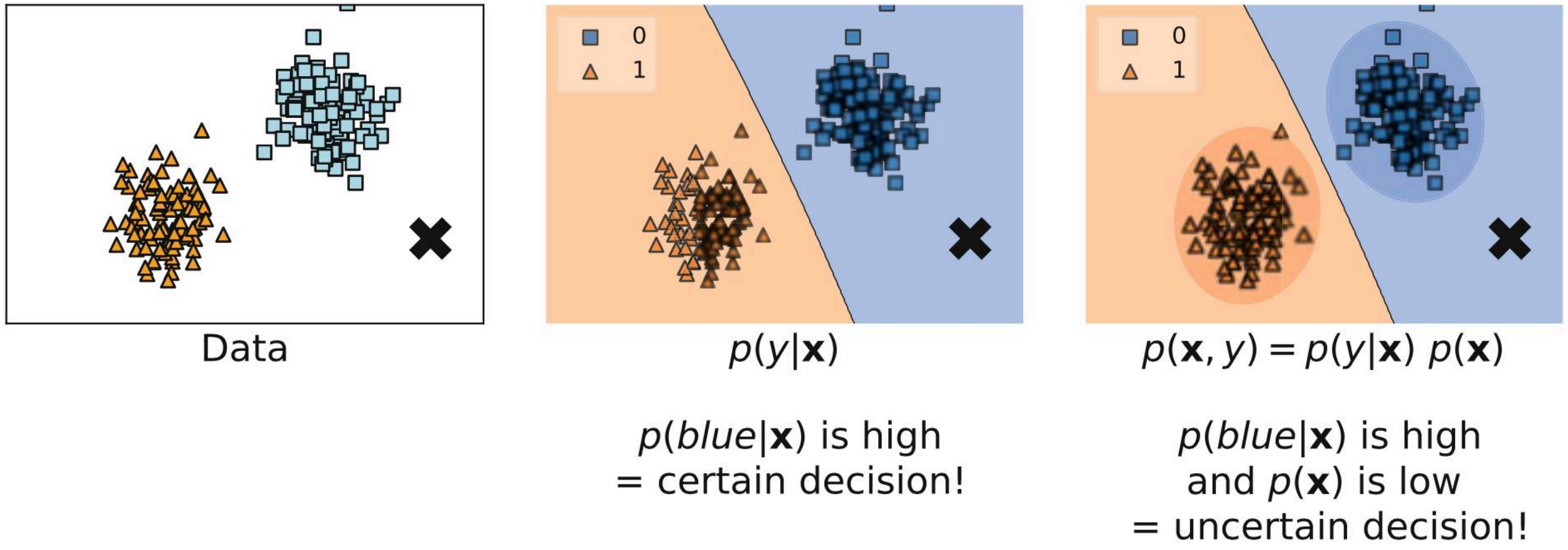


Fig. 1.2 And example of data (*left*) and two approaches to decision making: (*middle*) a discriminative approach and (*right*) a generative approach