CS542
Neural Computation With
Artificial Neural Networks

Lecture 04
Maximum Likelihood and
Graphical Models
Goal: Introduction to Probabilistic Models and Graphical Models
  - Maximum likelihood estimation
  - Notation of graphical models
  - Examples of graphical models and the implications derived from the graph

Handout:
  - Class Notes

Reading Assignment for Next Class
  - Bishop, Chapter 3
The View of Graphical Models

- They provide as simple way to visualize the structure of a probabilistic model and can be used to design and motivate new models.
- Insight into the properties of the model, including conditional independence, can be obtained by inspection of the graph.
- Complex computations, required to perform inference and learning in sophisticated models, can be expressed in terms of graphics manipulations, in which underlying mathematical expressions are carried along implicitly.
- Note that graphical models are useful for frequentist statistics as well as full Bayesian statistics.
Maximum Likelihood

One form of probabilistic derivation of learning algorithms

- Every data point and parameter is modeled by a probability distribution, e.g.,

\[ l = p(X,T) = \prod_{n} p(x^n,t^n) \]  (assume independent data)

\[ = \prod_{n} p(t^n | x^n)p(x^n) \]

- This “cost function” is maximized w.r.t. the open parameters in the probability functions
Maximum Likelihood
One form of probabilistic derivation of learning algorithms

- It is more convenient to optimize the log likelihood due to its simpler algebraic form, e.g.,

\[
\log l = \log p(X, T) \\
= \log \prod_n p(x^n, t^n) \quad \text{(assume independent data)} \\
= \log \prod_n p(t^n | x^n) p(x^n) \\
= \sum_n \log p(t^n | x^n) + \sum_n \log p(x^n)
\]
Maximum Likelihood Example

- Motivating the least squares criterion in regression
Maximum Likelihood Example

\[ p(t|x, w, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n | y(x_n, w), \beta^{-1}) \]

\[
\ln p(t|x, w, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, w) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) \\ \underbrace{\beta E(w)}
\]

Determine \( w_{ML} \) by minimizing sum-of-squares error,

\[
\frac{1}{\beta_{ML}} = \frac{1}{N} \sum_{n=1}^{N} \{y(x_n, w_{ML}) - t_n\}^2
\]
Bayesian Modeling

- In Bayesian modeling, parameters are not just treated as “point estimates”, but rather as complete distributions.
- Example: Linear regression:

\[
p(t, w | x, \beta) = N(w | w_0, \sigma_w^2) \prod_{N=1}^{N} N(t_n | y(x_n, w), \beta^{-1})
\]
General Probabilistic Modeling: Bayesian Networks

- Directed Acyclic Graph (DAG)

\[ p(a, b, c) = p(c|a, b)p(a, b) = p(c|a, b)p(b|a)p(a) \]

\[ p(x_1, \ldots, x_K) = p(x_K|x_1, \ldots, x_{K-1}) \cdots p(x_2|x_1)p(x_1) \]
Bayesian Networks

\[ p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \]
\[ p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5) \]

General Factorization

\[ p(x) = \prod_{k=1}^{K} p(x_k|pa_k) \]
Example: Polynomial Regression (1)

\[ y(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j \]

\[ p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n | y(\mathbf{w}, x_n)) \]
Example: Polynomial Regression (2)

\[ p(t, w) = p(w) \prod_{n=1}^{N} p(t_n | y(w, x_n)) \]
Example: Polynomial Regression (3)

- Input variables and explicit hyperparameters

\[ p(t, w | x, \alpha, \sigma^2) = p(w | \alpha) \prod_{n=1}^{N} p(t_n | w, x_n, \sigma^2). \]
Example: Polynomial Regression – Learning

- Condition on data

\[ p(w|t) \propto p(w) \prod_{n=1}^{N} p(t_n|w) \]
Example: Polynomial Regression — Prediction

Predictive distribution: \( p(\hat{t}|\hat{x}, x, t, \alpha, \sigma^2) \propto \int p(\hat{t}, t, w|\hat{x}, x, \alpha, \sigma^2) \, dw \)

where

\[
p(\hat{t}, t, w|\hat{x}, x, \alpha, \sigma^2) = \left[ \prod_{n=1}^{N} p(t_n|x_n, w, \sigma^2) \right] p(w|\alpha)p(\hat{t}|\hat{x}, w, \sigma^2)
\]
Generative Models

- E.g., causal process for generating images

![Diagram showing a causal process with nodes labeled Object, Position, Orientation, and Image.]
Discrete Variables (1)

- General joint distribution: \( K^2 - 1 \) parameters

\[
p(x_1, x_2 | \mu) = \prod_{k=1}^{K} \prod_{l=1}^{K} \mu_{kl}^{x_{1k} x_{2l}}
\]

- Independent joint distribution: \( 2(K - 1) \) parameters

\[
\hat{p}(x_1, x_2 | \mu) = \prod_{k=1}^{K} \mu_{1k}^{x_{1k}} \prod_{l=1}^{K} \mu_{2l}^{x_{2l}}
\]
Discrete Variables (2)

General joint distribution over $M$ variables: $K^M - 1$ parameters

$M$-node Markov chain: $K - 1 + (M - 1)K(K - 1)$ parameters
Discrete Variables: Bayesian Parameters (1)

\[
p(\{x_m, \mu_m\}) = p(x_1 | \mu_1) p(\mu_1) \prod_{m=2}^{M} p(x_m | x_{m-1}, \mu_m) p(\mu_m)
\]

\[
p(\mu_m) = \text{Dir}(\mu_m | \alpha_m)
\]
Discrete Variables: Bayesian Parameters (2)

\[ p \left( \{x_m\}, \mu_1, \mu \right) = p(x_1 | \mu_1) p(\mu_1) \prod_{m=2}^{M} p(x_m | x_{m-1}, \mu) p(\mu) \]
Parameterized Conditional Distributions

If $x_1, \ldots, x_M$ are discrete, K-state variables, $p(y = 1|x_1, \ldots, x_M)$ in general has $O(K^M)$ parameters.

The parameterized form

$$p(y = 1|x_1, \ldots, x_M) = \sigma \left( w_0 + \sum_{i=1}^{M} w_i x_i \right) = \sigma(w^T x)$$

requires only $M + 1$ parameters.
Linear-Gaussian Models

- Directed Graph

\[ p(x_i | pa_i) = \mathcal{N}\left(x_i \left| \sum_{j \in pa_i} w_{ij} x_j + b_i, \nu_i \right. \right) \]

Each node is Gaussian, the mean is a linear function of the parents.

- Vector-valued Gaussian Nodes

\[ p(x_i | pa_i) = \mathcal{N}\left(x_i \left| \sum_{j \in pa_i} W_{ij} x_j + b_i, \Sigma_i \right. \right) \]
Conditional Independence

• a is independent of b given c

\[ p(a|b, c) = p(a|c) \]

• Equivalently

\[ p(a, b|c) = p(a|b, c)p(b|c) = p(a|c)p(b|c) \]

• Notation

\[ a \perp b \mid c \]
Conditional Independence: Example 1

\[ p(a, b, c) = p(a|c)p(b|c)p(c) \]

\[ p(a, b) = \sum_c p(a|c)p(b|c)p(c) \]

\[ a \perp b \mid \emptyset \]
Conditional Independence: Example 1

\[
p(a, b | c) = \frac{p(a, b, c)}{p(c)} = p(a | c)p(b | c)
\]

\[a \indep b \mid c\]
Conditional Independence: Example 2

\[
p(a, b, c) = p(a)p(c|a)p(b|c)
\]

\[
p(a, b) = p(a) \sum_c p(c|a)p(b|c) = p(a)p(b|a)
\]

\[a \perp b \mid \emptyset\]
Conditional Independence: Example 2

\[ p(a, b|c) = \frac{p(a, b, c)}{p(c)} \]
\[ = \frac{p(a)p(c|a)p(b|c)}{p(c)} \]
\[ = p(a|c)p(b|c) \]

\[ a \perp b \mid c \]
Conditional Independence: Example 3

Note: this is the opposite of Example 1, with c unobserved.

\[ p(a, b, c) = p(a)p(b)p(c | a, b) \]

\[ p(a, b) = p(a)p(b) \]

\[ a \perp b \mid \emptyset \]
Conditional Independence: Example 3

Note: this is the opposite of Example 1, with $c$ observed.
“Am I out of fuel?”

\[
\begin{align*}
p(G = 1|B = 1, F = 1) &= 0.8 \\
p(G = 1|B = 1, F = 0) &= 0.2 \\
p(G = 1|B = 0, F = 1) &= 0.2 \\
p(G = 1|B = 0, F = 0) &= 0.1 \\
p(B = 1) &= 0.9 \\
p(F = 1) &= 0.9 \\
\text{and hence} \\
p(F = 0) &= 0.1 \\
\end{align*}
\]

\begin{align*}
B &= \text{Battery (0=flat, 1=fully charged)} \\
F &= \text{Fuel Tank (0=empty, 1=full)} \\
G &= \text{Fuel Gauge Reading} \\
&\quad \text{(0=empty, 1=full)}
\end{align*}
“Am I out of fuel?”

\[
p(F = 0 | G = 0) = \frac{p(G = 0 | F = 0) p(F = 0)}{p(G = 0)} \\
\approx 0.257
\]

Probability of an empty tank increased by observing G = 0.
“Am I out of fuel?”

\[
p(F = 0|G = 0, B = 0) = \frac{p(G = 0|B = 0, F = 0)p(F = 0)}{\sum_{F \in \{0,1\}} p(G = 0|B = 0, F) p(F)} \\
\approx 0.111
\]

Probability of an empty tank reduced by observing \( B = 0 \).
This referred to as “explaining away”.
A Fun Example

- There are 3 doors
- Behind one is a treasure
- You get one chance to pick one door (but do not open it yet)
- The Quizmaster tells you which of the two doors you did not pick does NOT contain the treasure.
- THE BIG QUESTION: Do you stick to your original door or do you switch to the other one?