

DSC291: Machine Learning with Few Labels

Machine Learning Basics

Zhiting Hu

Lecture 2, January 11, 2023

Lecture 3, January 13, 2023

UC San Diego

HALICIOĞLU DATA SCIENCE INSTITUTE

Outline

- Probability
 - Bayes' rule
 - Exponential family
 - KL divergence, cross entropy
- Functional derivatives
- Practice: MLE vs Maximum entropy

Probability

Why Probability?

- The world is a very uncertain place
 - “What will the weather be like today?”
 - “Will I like this movie?”
- We often can't prove something is true, but we can still ask how likely different outcomes are or ask for the most likely explanations
- Predictions need to have associated confidence
 - Confidence -> probability
- Not all machine learning models are probabilistic
 - ... but most of them have probabilistic interpretations



Notations

- A random variable \mathbf{x} represents outcomes or states of the world.
 - We write $p(\mathbf{x}_0)$ to mean Probability($\mathbf{x} = \mathbf{x}_0$)
- Sample space: the space of all possible outcomes (may be discrete, continuous, or mixed)
- $p(\mathbf{x})$ is the probability mass (density) function
 - Assigns a number to each point in sample space
 - Non-negative, sums (integrates) to 1
 - Intuitively: how often does \mathbf{x} occur, how much do we believe in \mathbf{x} .

Notations

- Joint distribution $p(\mathbf{x}, \mathbf{y})$
- Conditional distribution $p(\mathbf{y}|\mathbf{x})$

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

- Expectation:

$$\mathbb{E}[f(\mathbf{x})] = \sum_{\mathbf{x}} f(\mathbf{x}) p(\mathbf{x})$$

or

$$\mathbb{E}[f(\mathbf{x})] = \int_{\mathbf{x}} f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

Rules of Probability

- Sum rule

$$p(x) = \sum_y p(x, y) \quad (\text{Marginalize out } y)$$

$$p(x_1) = \sum_{x_2} \sum_{x_3} \dots \sum_{x_N} p(x_1, x_2, \dots, x_N)$$

- Product/chain rule

$$p(x, y) = p(y | x) p(x)$$

$$p(x_1, \dots, x_N) = p(x_1) p(x_2 | x_1) \dots p(x_N | x_1, \dots, x_{N-1})$$

Bayes' Rule

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

- This gives us a way of “reversing” conditional probabilities
- We call $p(\mathbf{y})$ the “prior”, and $p(\mathbf{y}|\mathbf{x})$ the “posterior”
- Ex: Bayes' Rule in machine learning:
 - \mathcal{D} : data (evidence)
 - θ : unknown quantities, such as model parameters, predictions

Posterior belief on the unknown quantities you see data \mathcal{D}

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$

Likelihood: How likely is the observed data under the particular unknown quantities θ

Prior belief on the unknown quantities **Before** you see data \mathcal{D}

Independence

- Two random variables are said to be **independent** iff their joint distribution factors

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$$

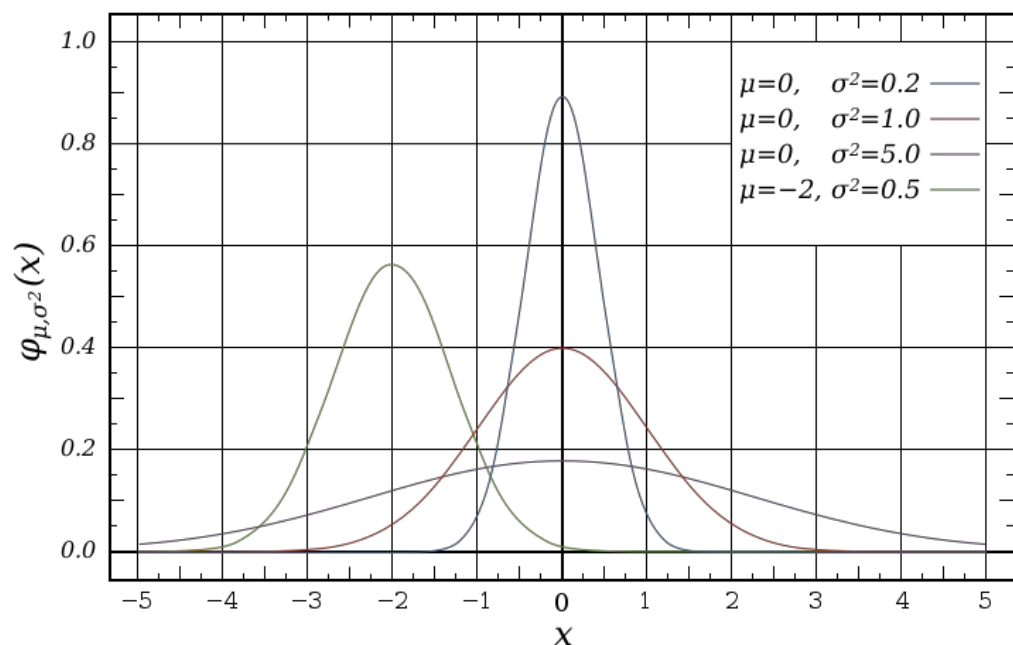
- Two random variables are **conditionally independent** given a third if they are independent after conditioning on the third

$$p(\mathbf{x}, \mathbf{y}|\mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z})$$

Some common distributions - Gaussian distribution

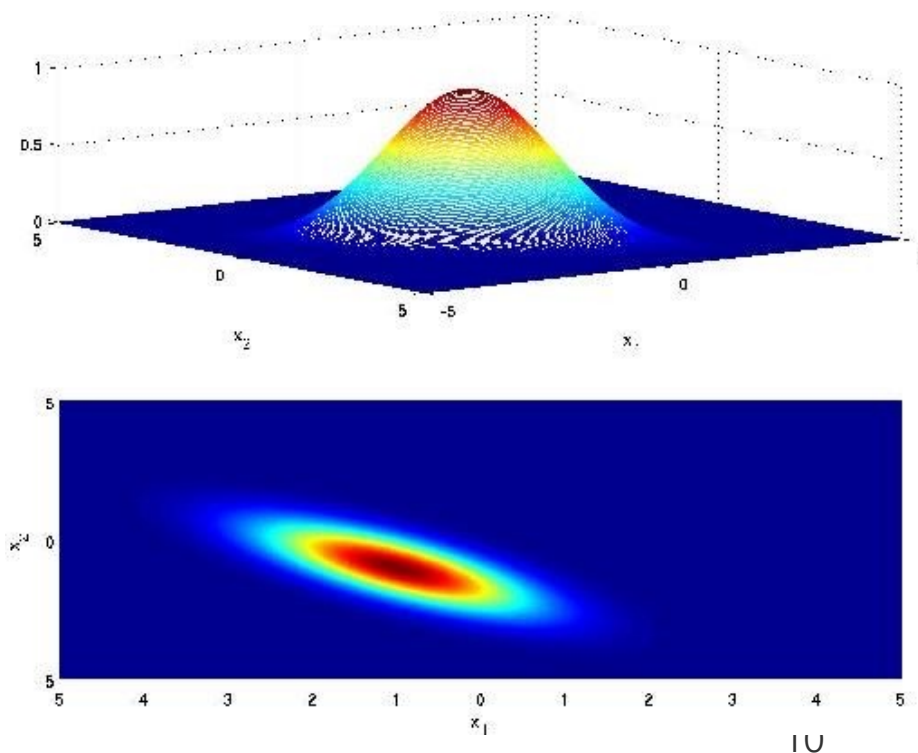
- Gaussian distribution

$$P(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$$



(Multivariate)

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





Some common distributions - Multinomial distribution

- Multinomial distribution
 - Discrete random variable \mathbf{x} that takes one of M values $\{1, \dots, M\}$
 - $p(\mathbf{x} = i) = \pi_i, \quad \sum_i \pi_i = 1$
 - Out of n independent trials, let k_i be the number of times $\mathbf{x} = i$ was observed
 - The probability of observing a vector of occurrences $\mathbf{k} = [k_1, \dots, k_M]$ is given by the *multinomial distribution* parametrized by $\boldsymbol{\pi}$

$$p(\mathbf{k}|\boldsymbol{\pi}, n) = p(k_1, \dots, k_m | \pi_1, \dots, \pi_m, n) = \frac{n!}{k_1! k_2! \dots k_m!} \prod_{i=1} \pi_i^{k_i}$$

- E.g., describing a text document by the frequency of occurrence of every distinct word
- For $n = 1$, a.k.a. *categorical distribution*
 - $p(\mathbf{x} = i | \boldsymbol{\pi}) = \pi_i$
 - In $\mathbf{k} = [k_1, \dots, k_M]$: $k_i = 1$, and $k_j = 0$ for all $j \neq i \rightarrow$ a.k.a., *one-hot representation* of i

Exponential family

- A distribution

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = h(\mathbf{x}) \exp\{\boldsymbol{\theta} \cdot T(\mathbf{x})\} / Z(\boldsymbol{\theta})$$

is an exponential family distribution

- $\boldsymbol{\theta} \in R^d$: natural (canonical) parameter
 - $T(\mathbf{x}) \in R^d$: sufficient statistics, features of data \mathbf{x}
 - $Z(\boldsymbol{\theta}) = \sum_{\mathbf{x}, y} h(\mathbf{x}) \exp\{\boldsymbol{\theta} \cdot T(\mathbf{x})\}$: normalization factor
- Examples: Bernoulli, multinomial, Gaussian, Poisson, gamma,...

Example: Multivariate Gaussian Distribution

- For a continuous vector random variable $\mathbf{x} \in R^k$

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

Moment parameter

$$= \frac{1}{(2\pi)^{k/2}} \exp\left\{-\frac{1}{2} \text{tr}(\Sigma^{-1} \mathbf{x} \mathbf{x}^T) + \mu^T \Sigma^{-1} \mathbf{x} - \frac{1}{2} \mu^T \Sigma^{-1} \mu - \log|\Sigma|\right\}$$

- Exponential family representation

$$\boldsymbol{\theta} = \left[\Sigma^{-1} \mu; -\frac{1}{2} \text{vec}(\Sigma^{-1}) \right] = [\boldsymbol{\theta}_1, \text{vec}(\boldsymbol{\theta}_2)], \quad \boldsymbol{\theta}_1 = \Sigma^{-1} \mu \text{ and } \boldsymbol{\theta}_2^- = -\frac{1}{2} \Sigma^{-1}$$

$$T(\mathbf{x}) = [\mathbf{x}; \text{vec}(\mathbf{x} \mathbf{x}^T)]$$

$$A(\boldsymbol{\theta}) = \frac{1}{2} \mu^T \Sigma^{-1} \mu + \log |\Sigma| = -\frac{1}{2} \text{tr}(\boldsymbol{\theta}_2 \boldsymbol{\theta}_1 \boldsymbol{\theta}_1^T) - \frac{1}{2} \log(-2\boldsymbol{\theta}_2)$$

$$h(\mathbf{x}) = (2\pi)^{-k/2}$$

Entropy

- Shannon entropy $H(p) = - \sum_x p(x) \log p(x)$
 - The average level of "information", "surprise", or "uncertainty" inherent to the variable x 's possible outcomes

KL Divergence

- Kullback-Leibler (KL) divergence: measures the closeness of two distributions $p(\mathbf{x})$ and $q(\mathbf{x})$

$$\text{KL}(q(\mathbf{x}) \parallel p(\mathbf{x})) = \sum_{\mathbf{x}} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})}$$

- a.k.a. Relative entropy
- $\text{KL} \geq 0$ (Jensen's inequality)
- Intuitively:
 - If q is high and p is high, then we are happy (i.e. low KL divergence)
 - If q is high and p is low then we pay a price (i.e. high KL divergence).
 - If q is low then we don't care (i.e. also low KL divergence, regardless of p)
- not a true "distance":
 - not commutative (symmetric) $\text{KL}(p \parallel q) \neq \text{KL}(q \parallel p)$
 - doesn't satisfy triangle inequality

KL Divergence

- Kullback-Leibler (KL) divergence: measures the closeness of two distributions $p(\mathbf{x})$ and $q(\mathbf{x})$

$$\text{KL}(q(\mathbf{x}) \parallel p(\mathbf{x})) = \sum_{\mathbf{x}} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})}$$

- a.k.a. Relative entropy
- Maximum likelihood estimation (MLE) is minimizing the KL divergence between the empirical data distribution and the model distribution

$$\text{KL}(\tilde{p}(\mathbf{x}) \parallel p_{\theta}(\mathbf{x})) = -\mathbb{E}_{\tilde{p}(\mathbf{x})}[\log p_{\theta}(\mathbf{x})] + H(\tilde{p}(\mathbf{x}))$$



Cross entropy

Key Takeaways

- Probability $p(\mathbf{x})$

- Bayes' rule $p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})}{p(\mathbf{x})}$
 - prior, posterior

- Exponential family:
 - Gaussian, multinomial, categorical, ...

$$\text{KL}(q(\mathbf{x}) || p(\mathbf{x})) = \sum_{\mathbf{x}} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})}$$

- KL Divergence
 - relation to Cross-entropy

Functional Derivatives

Functional derivative

- $\nabla_q - \mathbb{H}(q) = \log q + 1$
- Functional $F(y)$: an operator that takes a function $y(x)$ and returns an output value F
- Functional derivative (aka, variational derivative): relates a change in a Functional $F(y)$ to a change in the function y

Functional derivative

- Recall the conventional derivative $\frac{dy}{dx}$
 - Taylor expansion

$$y(x + \epsilon) = y(x) + \frac{dy}{dx}\epsilon + O(\epsilon^2)$$

- Functional derivative
 - How much a functional $F[y]$ changes when we make a small change $\epsilon\eta(x)$ to the function $y(x)$

$$F[y(x) + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\delta F}{\delta y(x)} \eta(x) dx + O(\epsilon^2)$$

- A function $y(x)$ that maximizes (or minimizes) a functional $F[y]$ must satisfy

$$\frac{\delta F}{\delta y(x)} = 0 \text{ for all } x$$

Functional derivative

$$F[y(x) + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\delta F}{\delta y(x)} \eta(x) dx + O(\epsilon^2)$$

- Consider a functional that is defined by an integral over a function $G(y, x)$

$$F[y] = \int G(y, x) dx$$

- Consider variations in the function $y(x)$,

$$F[y + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\partial G}{\partial y} \eta(x) dx + O(\epsilon^2)$$

Functional derivative

$$F[y(x) + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\delta F}{\delta y(x)} \eta(x) dx + O(\epsilon^2)$$

- Consider a functional that is defined by an integral over a function $G(y, x)$

$$F[y] = \int G(y, x) dx$$

- Ex.1, $-\mathbb{H}(q) = \int q(x) \log q(x) dx$
 - $G = q(x) \log q(x)$
- Consider variations in the function $y(x)$,

$$F[y + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\partial G}{\partial y} \eta(x) dx + O(\epsilon^2)$$

Practice: Maximum likelihood vs Maximum Entropy

Supervised Maximum Likelihood

- Model to be learned $p_{\theta}(\mathbf{x})$
- Observe full data $\mathcal{D} = \{ \mathbf{x}^* \}$
 - i.i.d: independent, identically distributed
- Maximum Likelihood Estimation (MLE)
 - The most classical learning algorithm

$$\min_{\theta} - \mathbb{E}_{\mathbf{x}^* \sim \mathcal{D}} \left[\log p_{\theta}(\mathbf{x}^*) \right]$$

- MLE is closely connected to the Maximum Entropy (MaxEnt) principle

Recap: Exponential Family

- A distribution

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = h(\mathbf{x}) \exp\{\boldsymbol{\theta} \cdot T(\mathbf{x})\} / Z(\boldsymbol{\theta})$$

is an exponential family distribution

- $\boldsymbol{\theta} \in R^d$: natural (canonical) parameter
 - $T(\mathbf{x}) \in R^d$: sufficient statistics, features of data \mathbf{x}
 - $Z(\boldsymbol{\theta}) = \sum_{\mathbf{x}, y} h(\mathbf{x}) \exp\{\boldsymbol{\theta} \cdot T(\mathbf{x})\}$: normalization factor
- Examples: Bernoulli, multinomial, Gaussian, Poisson, gamma,...

Maximum Likelihood for Exponential Family

$m(\mathbf{x})$: the number of times \mathbf{x} is observed in \mathcal{D}

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) &= \sum_{\mathbf{x}} m(\mathbf{x}) \log p(\mathbf{x} | \boldsymbol{\theta}) \\ &= \sum_{\mathbf{x}} m(\mathbf{x}) \left(\sum_i \theta_i T_i(\mathbf{x}) - \log Z(\boldsymbol{\theta}) \right) \\ &= \sum_{\mathbf{x}} m(\mathbf{x}) \sum_i \theta_i T_i(\mathbf{x}) - N \log Z(\boldsymbol{\theta})\end{aligned}$$

- Take gradient and set to 0

$$\begin{aligned}\frac{\partial}{\partial \theta_i} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) &= \sum_{\mathbf{x}} m(\mathbf{x}) T_i(\mathbf{x}) - N \frac{\partial}{\partial \theta_i} \log Z(\boldsymbol{\theta}) \\ &= \sum_{\mathbf{x}} m(\mathbf{x}) T_i(\mathbf{x}) - N \sum_{\mathbf{x}} p(\mathbf{x} | \boldsymbol{\theta}) T_i(\mathbf{x})\end{aligned}$$

$$\Rightarrow \boxed{\sum_{\mathbf{x}} p(\mathbf{x} | \boldsymbol{\theta}) T_i(\mathbf{x})} = \sum_{\mathbf{x}} \frac{m(\mathbf{x})}{N} T_i(\mathbf{x}) = \boxed{\sum_{\mathbf{x}} \tilde{p}(\mathbf{x} | \boldsymbol{\theta}) T_i(\mathbf{x})}$$

At MLE, the expectations of the sufficient statistics under the model must match empirical feature average

Maximum Entropy (MaxEnt)

- Given \mathcal{D} , to estimate $p(\mathbf{x})$
- We can approach the problem from an entirely different point of view. Begin with some fixed feature expectations:

$$\sum_{\mathbf{x}} p(\mathbf{x}) T_i(\mathbf{x}) = \sum_{\mathbf{x}} \frac{m(\mathbf{x})}{N} T_i(\mathbf{x}) := \alpha_i$$

- There may exist many distributions which satisfy them. Which one should we select?
 - MaxEnt principle: the most uncertain or flexible one, i.e., the one with maximum entropy
- This yields a new optimization problem:
 - This is a variational definition of a distribution!

$$\max_p \quad \mathbb{H}(p(\mathbf{x})) = - \sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x})$$

$$\text{s.t.} \quad \sum_{\mathbf{x}} p(\mathbf{x}) T_i(\mathbf{x}) = \alpha_i$$

$$\sum_{\mathbf{x}} p(\mathbf{x}) = 1$$

Solution to the MaxEnt Problem

- To solve the MaxEnt problem, we use Lagrange multipliers:

$$\max_{\theta, \mu} \min_{p(\mathbf{x})} L = - \sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x}) - \sum_i \theta_i \left(\sum_{\mathbf{x}} p(\mathbf{x}) T_i(\mathbf{x}) - \alpha_i \right) - \mu \left(\sum_{\mathbf{x}} p(\mathbf{x}) - 1 \right)$$

Solution to the MaxEnt Problem

- To solve the MaxEnt problem, we use Lagrange multipliers:

$$\max_{\theta, \mu} \min_{p(\mathbf{x})} L = - \sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x}) - \sum_i \theta_i \left(\sum_{\mathbf{x}} p(\mathbf{x}) T_i(\mathbf{x}) - \alpha_i \right) - \mu \left(\sum_{\mathbf{x}} p(\mathbf{x}) - 1 \right)$$

$$\frac{\partial L}{\partial p(\mathbf{x})} = 1 + \log p(\mathbf{x}) - \sum_i \theta_i T_i(\mathbf{x}) - \mu$$

$$p^*(\mathbf{x}) = e^{\mu-1} \exp \left\{ \sum_i \theta_i f_i(\mathbf{x}) \right\}$$

$$Z(\boldsymbol{\theta}) = e^{\mu-1} = \sum_{\mathbf{x}} \exp \left\{ \sum_i \theta_i f_i(\mathbf{x}) \right\} \quad \left(\text{since } \sum_{\mathbf{x}} p^*(\mathbf{x}) = 1 \right)$$

$$p(\mathbf{x} | \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp \left\{ \sum_i \theta_i T_i(\mathbf{x}) \right\}$$

Solution to the MaxEnt Problem

- To solve the MaxEnt problem, we use Lagrange multipliers:

$$\max_{\theta, \mu} \min_{p(\mathbf{x})} L = - \sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x}) - \sum_i \theta_i \left(\sum_{\mathbf{x}} p(\mathbf{x}) T_i(\mathbf{x}) - \alpha_i \right) - \mu \left(\sum_{\mathbf{x}} p(\mathbf{x}) - 1 \right)$$

$$\frac{\partial L}{\partial p(\mathbf{x})} = 1 + \log p(\mathbf{x}) - \sum_i \theta_i T_i(\mathbf{x}) - \mu$$

$$p^*(\mathbf{x}) = e^{\mu-1} \exp \left\{ \sum_i \theta_i f_i(\mathbf{x}) \right\}$$

$$Z(\boldsymbol{\theta}) = e^{\mu-1} = \sum_{\mathbf{x}} \exp \left\{ \sum_i \theta_i f_i(\mathbf{x}) \right\} \quad \left(\text{since } \sum_{\mathbf{x}} p^*(\mathbf{x}) = 1 \right)$$

$$p(\mathbf{x} | \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp \left\{ \sum_i \theta_i T_i(\mathbf{x}) \right\}$$

- So feature constraints + MaxEnt \Rightarrow **exponential family**.
- Problem is strictly convex w.r.t. $p(\mathbf{x})$, so solution is unique.

Solution to the MaxEnt Problem

- To solve the MaxEnt problem, we use Lagrange multipliers:

$$\max_{\theta, \mu} \min_{p(\mathbf{x})} L = - \sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x}) - \sum_i \theta_i \left(\sum_{\mathbf{x}} p(\mathbf{x}) T_i(\mathbf{x}) - \alpha_i \right) - \mu \left(\sum_{\mathbf{x}} p(\mathbf{x}) - 1 \right)$$

$$p(\mathbf{x} | \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp \left\{ \sum_i \theta_i T_i(\mathbf{x}) \right\}$$

plug $p(\mathbf{x} | \boldsymbol{\theta})$ back into L , and since $\sum_{\mathbf{x}} \frac{m(\mathbf{x})}{N} T_i(\mathbf{x}) := \alpha_i$:

$$\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \sum_{\mathbf{x}} m(\mathbf{x}) \sum_i \theta_i T_i(\mathbf{x}) - N \log Z(\boldsymbol{\theta})$$

- Recovers precisely the MLE problem of exponential family

- So feature constraints + MaxEnt \Rightarrow **exponential family**.
- Problem is strictly convex w.r.t. $p(\mathbf{x})$, so solution is unique.

Constraints from Data

- We have seen a case of **convex duality**:
 - In one case, we assume exponential family and show that Maximum Likelihood implies model expectations must match empirical expectations.
 - In the other case, we assume model expectations must match empirical feature counts and show that MaxEnt implies exponential family distribution.

A more general MaxEnt problem

$$\min_p \text{KL}(p(\mathbf{x}) \| h(\mathbf{x}))$$

$$\stackrel{\text{def}}{=} \sum_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{h(\mathbf{x})} = -\text{H}(p) - \sum_{\mathbf{x}} p(\mathbf{x}) \log h(\mathbf{x})$$

$$\text{s.t.} \quad \sum_{\mathbf{x}} p(\mathbf{x}) T_i(\mathbf{x}) = \alpha_i$$

$$\sum_{\mathbf{x}} p(\mathbf{x}) = 1$$

$$\Rightarrow p(\mathbf{x} | \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp \left\{ \sum_i \theta_i T_i(\mathbf{x}) \right\}$$

Summary

- Maximum entropy is dual to maximum likelihood of exponential family distributions
- This provides an alternative view of the problem of fitting a model into data:
 - The data instances in the training set are treated as constraints, and the learning problem is treated as a constrained optimization problem.
 - We'll revisit this optimization-theoretic view of learning repeatedly in the future!

$$\begin{aligned} \max_p \quad & H(p(\mathbf{x})) = - \sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x}) \\ \text{s.t.} \quad & \sum_{\mathbf{x}} p(\mathbf{x}) T_i(\mathbf{x}) = \alpha_i \\ & \sum_{\mathbf{x}} p(\mathbf{x}) = 1 \end{aligned}$$

Key Takeaways

- Probability
 - Bayes' rule
 - Exponential family
 - KL divergence
- Functional derivative
- Convex duality between MLE and MaxEnt (optional)

Questions?