Speech Recognition

Lecture 7: Maximum Entropy Models

Mehryar Mohri
Courant Institute and Google Research
mohri@cims.nyu.com

This Lecture

- Information theory basics
- Maximum entropy models
- Duality theorem

Convexity

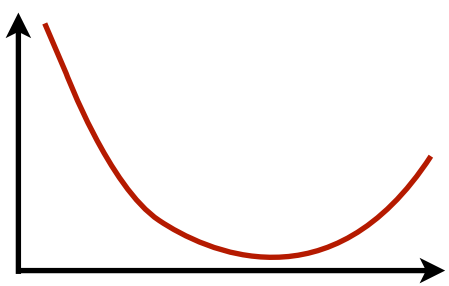
- **Definition**: $X \subseteq \mathbb{R}^N$ is said to be convex if for any two points $x, y \in X$ the segment $[x, y]$ lies in X:

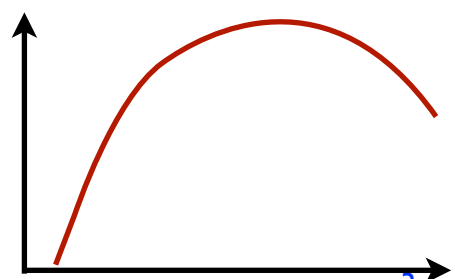
$$\{\alpha x + (1-\alpha)y, 0 \leq \alpha \leq 1\} \subseteq X.$$

- **Definition**: let X be a convex set. A function $f : X \to \mathbb{R}$ is said to be convex if for all $x, y \in X$ and $\alpha \in (0, 1)$,

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y).$$

With a strict inequality, f is said to be strictly convex. f is said to be concave when $-f$ is convex.

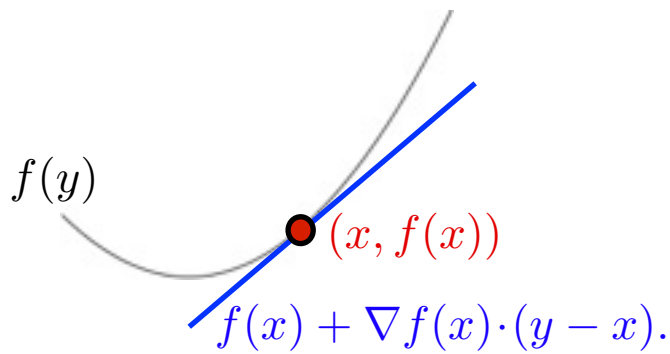




Properties of Convex Functions

- **Theorem**: let f be a differentiable function. Then, f is convex iff $\mathrm{dom}(f)$ is convex and

$$\forall x, y \in \mathrm{dom}(f),\ \ f(y) - f(x) \geq \nabla f(x) \cdot (y - x).$$



- **Theorem**: let f be a twice differentiable function. Then, f is convex iff its Hessian is positive semi-definite:

$$\forall x \in \mathrm{dom}(f),\ \ \nabla^2 f(x) \succeq 0.$$

Jensen's Inequality

- Theorem: let X be a random variable and f a measurable convex function. Then,

$$f(\mathrm{E}[X]) \leq \mathrm{E}[f(X)].$$

- Proof:
  - For a distribution over a finite set, the property follows directly the definition of convexity.
  - The general case is a consequence of the continuity of convex functions and the density of finite distributions.

Entropy

- **Definition**: the entropy of a random variable X with probability distribution $p(x) = \Pr[X = x]$ is

$$H(X) = -\mathrm{E}[\log p(X)] = -\sum_{x \in X} p(x) \log p(x).$$

- **Properties**:
  - measure of uncertainty of $p(x)$.
  - $H(X) \geq 0$.
  - maximal for uniform distribution. For a finite support, by Jensen's inequality:

$$H(X) = \mathrm{E}\left[\log \frac{1}{p(X)}\right] \leq \log \mathrm{E}\left[\frac{1}{p(X)}\right] = \log N.$$

Relative Entropy

- Definition: the relative entropy (or Kullback-Leibler divergence) of two distributions p and q is

$$D(p \parallel q) = \mathrm{E}_p\left[\log \frac{p(X)}{q(X)}\right] = \sum_x p(x) \log \frac{p(x)}{q(x)},$$

with $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$.

- Properties:
  - assymetric measure of divergence between two distributions. It is convex in p and q .
  - $D(p \parallel q) \geq 0$ for all p and q .
  - $D(p \parallel q) = 0$ iff $p = q$.

Non-Negativity of Relative Entropy

- Treat separately the case $q(x) = 0$ for $x \in \mathrm{supp}(p)$, or use the same proof by extending domain of $\log$:

$$\begin{aligned} -D(p \parallel q) &= \mathrm{E}_p \left[\log \frac{q(X)}{p(X)} \right] \\ &\leq \log \mathrm{E}_p \left[\frac{q(X)}{p(X)} \right] \\ &= \log \sum_x q(x) = 0. \end{aligned}$$

This Lecture

- Information theory basics
- Maximum entropy models
- Duality theorem

Problem

- Data: sample S drawn i.i.d. from set X according to some distribution D,

$$x_1, \ldots, x_m \in X.$$

- Problem: find distribution p out of a set $\mathcal{P}$ that best estimates D.

Maximum Likelihood

- Likelihood: probability of observing sample under distribution $p \in \mathcal{P}$, which, given the independence assumption is

$$\Pr[S] = \prod_{i=1}^{m} p(x_i).$$

- Principle: select distribution maximizing likelihood,

$$p_\star = \operatorname*{argmax}_{p \in \mathcal{P}} \prod_{i=1}^{m} p(x_i),$$

or, equivalently $p_\star = \operatorname*{argmax}_{p \in \mathcal{P}} \sum_{i=1}^{m} \log p(x_i).$

Relative Entropy Formulation

- Empirical distribution: distribution $\widehat{p}$ corresponding to sample S.
- Lemma: $p_\star$ has maximum likelihood iff

$$p_\star = \operatorname*{argmin}_{p \in \mathcal{P}} D(\widehat{p} \,\|\, p).$$

- Proof:

$$\begin{aligned} D(\hat{p} \,\|\, p) &= \sum_x \hat{p}(x) \log \hat{p}(x) - \sum_x \hat{p}(x) \log p(x) \\ &= -H(\hat{p}) - \sum_x \frac{|x|_S}{m} \log p(x) \\ &= -H(\hat{p}) - \frac{1}{m} \log \prod_{x \text{ in } S} p(x)^{|x|_S} \\ &= -H(\hat{p}) - \frac{1}{m} \log \left(\Pr_{S \sim p^m} [S] \right). \end{aligned}$$

Problem

- Data: sample S drawn i.i.d. from set X according to some distribution D,

$$x_1, \ldots, x_m \in X.$$

- Features: associated to elements of X,

$$\Phi \colon X \to \mathbb{R}^N.$$

- Problem: how do we estimate distribution D? Uniform distribution u over X?

Features

- Examples:
  - *n*-grams, distance-*d* *n*-grams.
  - class-based n-grams, word triggers.
  - sentence length.
  - number and type of verbs.
  - various grammatical information (e.g., agreement, POS tags).
  - dialog-level information.

Hoeffding's Bounds

- Theorem: let $X_1, X_2, \ldots, X_m$ be a sequence of independent Bernoulli trials taking values in $[0, 1]$, then for all $\epsilon > 0$, the following inequalities hold for $\overline{X}_m = \frac{1}{m}\sum_{i=1}^{m} X_i$:

$$\Pr\left[\overline{X}_m - \mathrm{E}[\overline{X}_m] \geq \epsilon\right] \leq e^{-2m\epsilon^2}$$

$$\Pr\left[\overline{X}_m - \mathrm{E}[\overline{X}_m] \leq -\epsilon\right] \leq e^{-2m\epsilon^2}.$$

Maximum Entropy Principle

(E. T. Jaynes, 1957, 1983)

- For large m, empirical average is a good estimate of the expected values of the features:

$$\mathop{\mathrm{E}}_{x \sim D}[\Phi_j(x)] \approx \mathop{\mathrm{E}}_{x \sim \widehat{p}}[\Phi_j(x)], \quad j \in [1, N].$$

- Principle: find distribution that is closest to the uniform distribution u and that preserves the expected values of features.
- Closeness is measured using relative entropy.

Maximum Entropy Formulation

- Distributions: let $\mathcal{P}$ denote the set of distributions

$$\mathcal{P} = \Big\{ p \in \Delta : \mathop{\mathrm{E}}_{x \sim p}[\Phi(x)] = \mathop{\mathrm{E}}_{x \sim \widehat{p}}[\Phi(x)] \Big\}.$$

- Optimization problem: find distribution $p_\star$ verifying

$$p_\star = \mathop{\mathrm{argmin}}_{p \in \mathcal{P}} D(p \parallel u).$$

Relation with Entropy Maximization

- Relationship with entropy:

$$D(p \parallel u) = \sum_{x \in X} p(x) \log \frac{p(x)}{1/|X|}$$

$$= \log |X| + \sum_{x \in X} p(x) \log p(x) = \log |X| - H(p).$$

- Optimization problem:

$$\begin{aligned} \text{minimize} \quad & \sum_{x \in X} p(x) \log p(x) \\ \text{subject to} \quad & p(x) \geq 0, \forall x \in X, \\ & \sum_{x \in X} p(x) = 1, \\ & \sum_{x \in X} p(x) \Phi(x) = \frac{1}{m} \sum_{i=1}^{m} \Phi(x_i). \end{aligned}$$

Maximum Likelihood Gibbs Distrib.

- Gibbs distributions: set $\mathcal{Q}$ of distributions p defined by

$$p(x) = \frac{1}{Z} \exp\big(\mathbf{w} \cdot \Phi(x)\big) = \frac{1}{Z} \exp\Big(\sum_{j=1}^{N} w_j \cdot \Phi_j(x)\Big),$$

with $Z = \sum_x \exp\big(\mathbf{w} \cdot \Phi(x)\big)$.

- Maximum likelihood Gibbs distribution:

$$p_\star = \operatorname*{argmax}_{q \in \overline{\mathcal{Q}}} \sum_{i=1}^{m} \log q(x_i) = \operatorname*{argmin}_{q \in \bar{\mathcal{Q}}} D(\hat{p} \parallel q),$$

where $\overline{\mathcal{Q}}$ is the closure of $\mathcal{Q}$.

Duality Theorem

(Della Pietra et al., 1997)

- Theorem: assume that $D(\hat{p} \parallel u) < \infty$. Then, there exists a unique probability distribution $p_\star$ satisfying

1. $p_\star \in \mathcal{P} \cap \bar{\mathcal{Q}}$;
2. $D(p \parallel q) = D(p \parallel p_\star) + D(p_\star \parallel q)$ for any $p \in \mathcal{P}$ and $q \in \overline{\mathcal{Q}}$ (Pythagorean equality);
3. $p_\star = \operatorname*{argmin}_{q \in \bar{\mathcal{Q}}} D(\hat{p} \parallel q)$ (maximum likelihood);
4. $p_\star = \operatorname*{argmin}_{p \in \mathcal{P}} D(p \parallel u)$ (maximum entropy).

Each of these properties determines $p_\star$ uniquely.

Regularization

- Relaxation: sample size too small to impose equality constraints. Instead,
  - L_1 constraints:

$$\left| \mathop{\mathrm{E}}_{x \sim D}[\Phi_j(x)] - \mathop{\mathrm{E}}_{x \sim \widehat{p}}[\Phi_j(x)] \right| \leq \beta_j, \quad j \in [1, N].$$

  - L_2 constraints:

$$\left\| \mathop{\mathrm{E}}_{x \sim D}[\Phi(x)] - \mathop{\mathrm{E}}_{x \sim \widehat{p}}[\Phi(x)] \right\|_2 \leq \Lambda^2.$$

L_1 Maxent

- Optimization problem:

$$\operatorname*{arginf}_{\mathbf{w}\in\mathbb{R}^N} \sum_{j=1}^{N} \beta_j |\mathbf{w}_j| - \frac{1}{m} \sum_{i=1}^{m} \log p_{\mathbf{w}}[x_i],$$

with $p_{\mathbf{w}}[x] = \frac{1}{Z} \exp\left(\mathbf{w} \cdot \mathbf{\Phi}(x)\right)$.

- Bayesian interpretation: Laplacian prior.

$$p_{\mathbf{w}}[x] \rightarrow p(\mathbf{w})\, p_{\mathbf{w}}[x]$$

with $p(\mathbf{w}) = \prod_{j=1}^{N} \frac{\beta_j}{2} \exp(-\beta_j |w_j|)$.

L_2 Maxent

- Optimization problem:

$$\operatorname*{arginf}_{\mathbf{w}\in\mathbb{R}^N} \lambda\|\mathbf{w}\|^2 - \frac{1}{m}\sum_{i=1}^{m} \log p_{\mathbf{w}}[x_i].$$

with $p_{\mathbf{w}}[x] = \frac{1}{Z}\exp\left(\mathbf{w}\cdot\mathbf{\Phi}(x)\right)$.

- Bayesian interpretation: Gaussian prior.

$$p_{\mathbf{w}}[x] \rightarrow p(\mathbf{w})\, p_{\mathbf{w}}[x]$$

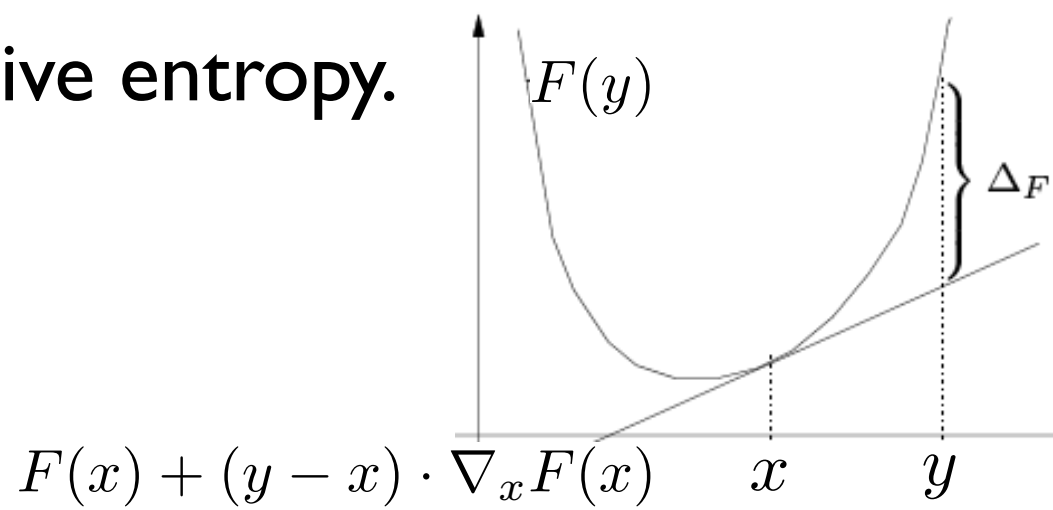
with $p(\mathbf{w}) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{w_j^2}{2\sigma^2}\right)$.

Extensions - Bregman Divergences

- Definition: let F be a convex and differentiable function, then the Bregman divergence based on F is defined as

$$B_F(y, x) = F(y) - F(x) - (y - x) \cdot \nabla_x F(x).$$

- Examples:
  - Unnormalized relative entropy.
  - Euclidean distance.



This Lecture

- Information theory basics
- Maximum entropy models
- Duality theorem

References

- Adam Berger, Stephen Della Pietra, and Vincent Della Pietra. A maximum entropy approach to natural language processing. Computational Linguistics, (22-1), March 1996;
- Berkson, J. (1944). Application of the logistic function to bio-assay. Journal of the American Statistical Association 39, 357–365.
- Imre Csiszar and Tusnady. Information geometry and alternating minimization procedures. Statistics and Decisions, Supplement Issue 1, 205-237, 1984.
- Imre Csiszar. A geometric interpretation of Darroch and Ratchliff's generalized iterative scaling. The Annals of Statisics, 17(3), pp. 1409-1413. 1989.
- J. Darroch and D. Ratchliff. Generalized iterative scaling for log-linear models. The Annals of Mathematical Statistics, 43(5), pp. 1470-1480, 1972.
- Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 19:4, pp.380--393, April, 1997.

References

- Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. *Duality and auxiliary functions for Bregman distances*. Technical Report CMU-CS-01-109, School of Computer Science, CMU, 2001.

- E. Jaynes. Information theory and statistical mechanics. Physics Reviews, 106:620–630, 1957.

- E. Jaynes. Papers on Probability, Statistics, and Statistical Physics. R. Rosenkrantz (editor), D. Reidel Publishing Company, 1983.

- O'Sullivan. Alternating minimzation algorithms: From Blahut-Arimoto to expectation-maximization. Codes, Curves and Signals: Common Threads in Communications, A. Vardy, (editor), Kluwer, 1998.

- Roni Rosenfeld. A maximum entropy approach to adaptive statistical language modelling. Computer, Speech and Language 10:187--228, 1996.