Foundations of Machine Learning

Maximum Entropy Models, Logistic Regression

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

Motivation

- Probabilistic models:
  - density estimation.
  - classification.

This Lecture

- Notions of information theory.
- Introduction to density estimation.
- Maxent models.
- Conditional Maxent models.

Entropy

(Shannon, 1948)

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

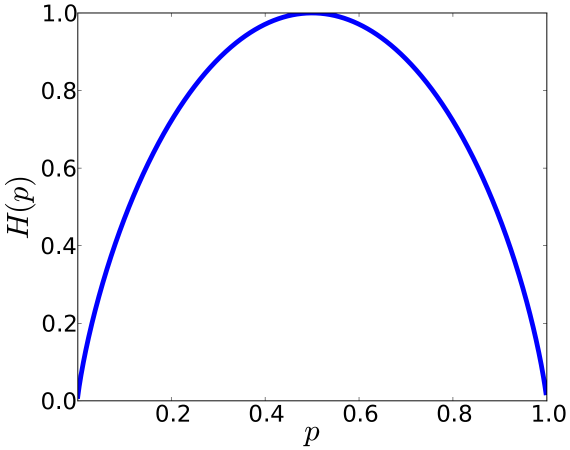
$$H(X) = -\mathrm{E}[\log \mathsf{p}(X)] = -\sum_{x \in X} \mathsf{p}(x) \log \mathsf{p}(x).$$

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

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

Entropy

- Base of logarithm: not critical; for base 2, $-\log_2(\mathsf{p}(x))$ is the number of bits needed to represent $\mathsf{p}(x)$.
- Definition and notation: the entropy of a distribution p is defined by the same quantity and denoted by $H(\mathsf{p})$.
- Special case of Rényi entropy (Rényi, 1961).
- Binary entropy: $H(\mathsf{p}) = -\mathsf{p}\log\mathsf{p} - (1-\mathsf{p})\log(1-\mathsf{p})$.



Relative Entropy

(Shannon, 1948; Kullback and Leibler, 1951)

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

$$D(\mathsf{p} \,\|\, \mathsf{q}) = \mathrm{E}_{\mathsf{p}}\left[\log \frac{\mathsf{p}(X)}{\mathsf{q}(X)}\right] = \sum_{x \in \mathcal{X}} \mathsf{p}(x) \log \frac{\mathsf{p}(x)}{\mathsf{q}(x)},$$

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

- Properties:
  - asymmetric: in general, $D(\mathsf{p} \,\|\, \mathsf{q}) \neq D(\mathsf{q} \,\|\, \mathsf{p})$ for $\mathsf{p} \neq \mathsf{q}$.
  - non-negative: $D(\mathsf{p} \,\|\, \mathsf{q}) \geq 0$ for all p and q.
  - definite: $(D(\mathsf{p} \,\|\, \mathsf{q}) = 0) \Rightarrow (\mathsf{p} = \mathsf{q})$.

Non-Negativity of Rel. Entropy

- By the concavity of log and Jensen's inequality,

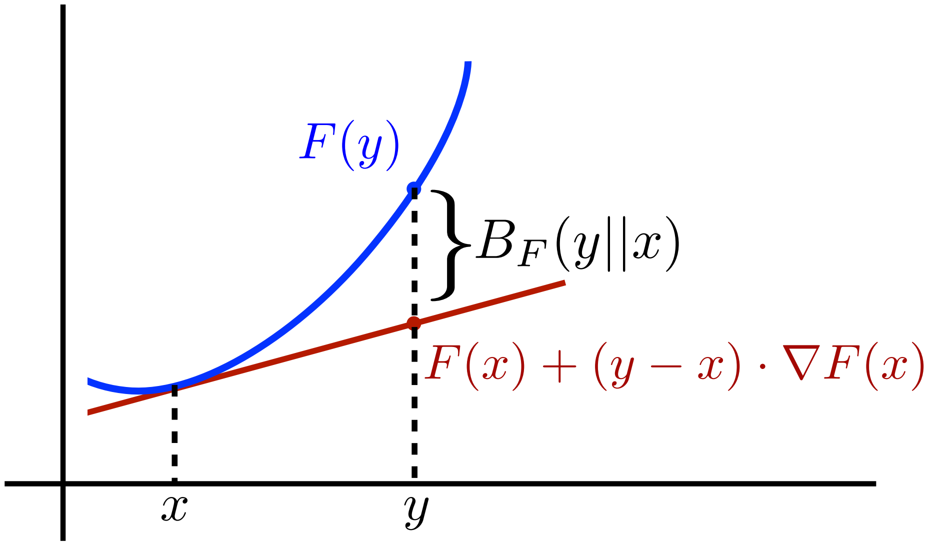
$$\begin{aligned}
-D(\mathsf{p} \parallel \mathsf{q}) &= \sum_{x\colon\, \mathsf{p}(x)>0} \mathsf{p}(x) \log\left(\frac{\mathsf{q}(x)}{\mathsf{p}(x)}\right) \\
&\leq \log\left(\sum_{x\colon\, \mathsf{p}(x)>0} \mathsf{p}(x)\frac{\mathsf{q}(x)}{\mathsf{p}(x)}\right) \\
&= \log\left(\sum_{x\colon\, \mathsf{p}(x)>0} \mathsf{q}(x)\right) \leq \log(1) = 0.
\end{aligned}$$

Bregman Divergence

(Bregman, 1967)

- Definition: let F be a convex and differentiable function defined over a convex set C in a Hilbert space $\mathbb{H}$. Then, the Bregman divergence B_F associated to F is defined by

$$B_F(x \parallel y) = F(x) - F(y) - \langle \nabla F(y), x - y \rangle \, .$$



Bregman Divergence

- Examples:

	$\boldsymbol{B_F(x \parallel y)}$	$\boldsymbol{F(x)}$
Squared L_2-distance	$\lVert\mathbf{x}-\mathbf{y}\rVert^2$	$\lVert\mathbf{x}\rVert^2$
Mahalanobis distance	$(\mathbf{x}-\mathbf{y})^\top\mathbf{K}^{-1}(\mathbf{x}-\mathbf{y})$	$\mathbf{x}^\top\mathbf{K}^{-1}\mathbf{x}$
Unnormalized relative entropy	$\widetilde{D}(\mathbf{x} \parallel \mathbf{y})$	$\sum_{i\in I} x_i \log x_i - x_i$

  - note: relative entropy not a Bregman divergence since not defined over an open set; but, on the simplex, coincides with unnormalized relative entropy

$$\widetilde{D}(\mathsf{p} \parallel \mathsf{q}) = \sum_{x\in\mathcal{X}} \mathsf{p}(x) \log\left[\frac{\mathsf{p}(x)}{\mathsf{q}(x)}\right] + \big(\mathsf{q}(x) - \mathsf{p}(x)\big).$$

Conditional Relative Entropy

- Definition: let p and q be two probability distributions over $\mathcal{X} \times \mathcal{Y}$. Then, the conditional relative entropy of p and q with respect to distribution r over $\mathcal{X}$ is defined by

$$\begin{aligned}\operatorname*{E}_{X \sim \mathsf{r}}\Big[D\big(\mathsf{p}(\cdot|X) \parallel \mathsf{q}(\cdot|X)\big)\Big] &= \sum_{x \in \mathcal{X}} \mathsf{r}(x) \sum_{y \in \mathcal{Y}} \mathsf{p}(y|x) \log \frac{\mathsf{p}(y|x)}{\mathsf{q}(y|x)} \\ &= D(\widetilde{\mathsf{p}} \parallel \widetilde{\mathsf{q}}),\end{aligned}$$

  with $\widetilde{\mathsf{p}}(x, y) = \mathsf{r}(x)\mathsf{p}(y|x)$, $\widetilde{\mathsf{q}}(x, y) = \mathsf{r}(x)\mathsf{q}(y|x)$, and the conventions $0 \log 0 = 0$, $0 \log \frac{0}{0} = 0$, and $\mathsf{p} \log \frac{\mathsf{p}}{0} = +\infty$.

  - note: the definition of conditional relative entropy is not intrinsic, it depends on a third distribution r.

This Lecture

- Notions of information theory.
- Introduction to density estimation.
- Maxent models.
- Conditional Maxent models.

Density Estimation Problem

- Training data: sample S of size m drawn i.i.d. from set $\mathcal{X}$ according to some distribution $\mathcal{D}$,

$$S = (x_1, \ldots, x_m).$$

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

Maximum Likelihood Solution

- Maximum Likelihood principle: select distribution $\mathsf{p} \in \mathcal{P}$ maximizing likelihood of observed sample S,

$$\begin{aligned} \mathsf{p}_{\mathrm{ML}} &= \operatorname*{argmax}_{\mathsf{p} \in \mathcal{P}} \Pr[S|\mathsf{p}] \\ &= \operatorname*{argmax}_{\mathsf{p} \in \mathcal{P}} \prod_{i=1}^{m} \mathsf{p}(x_i) \\ &= \operatorname*{argmax}_{\mathsf{p} \in \mathcal{P}} \sum_{i=1}^{m} \log \mathsf{p}(x_i). \end{aligned}$$

Relative Entropy Formulation

- Lemma: let $\widehat{\mathsf{p}}_S$ be the empirical distribution for sample S, then

$$\mathsf{p}_{\mathrm{ML}} = \underset{\mathsf{p} \in \mathcal{P}}{\operatorname{argmin}}\, D(\widehat{\mathsf{p}}_S \parallel \mathsf{p}).$$

- Proof:

$$\begin{aligned} D(\widehat{\mathsf{p}}_S \parallel \mathsf{p}) &= \sum_x \widehat{\mathsf{p}}_S(x) \log \widehat{\mathsf{p}}_S(x) - \sum_x \widehat{\mathsf{p}}_S(x) \log \mathsf{p}(x) \\ &= -H(\widehat{\mathsf{p}}_S) - \sum_x \frac{\sum_{i=1}^m 1_{x=x_i}}{m} \log \mathsf{p}(x) \\ &= -H(\widehat{\mathsf{p}}_S) - \sum_{i=1}^m \sum_x \frac{1_{x=x_i}}{m} \log \mathsf{p}(x) \\ &= -H(\widehat{\mathsf{p}}_S) - \sum_{i=1}^m \frac{\log \mathsf{p}(x_i)}{m}. \end{aligned}$$

Maximum a Posteriori (MAP)

- Maximum a Posteriori principle: select distribution $\mathsf{p} \in \mathcal{P}$ that is the most likely, given the observed sample S and assuming a prior distribution $\Pr[\mathsf{p}]$ over $\mathcal{P}$,

$$\begin{aligned} \mathsf{p}_{\text{MAP}} &= \operatorname*{argmax}_{\mathsf{p} \in \mathcal{P}} \Pr[\mathsf{p}|S] \\ &= \operatorname*{argmax}_{\mathsf{p} \in \mathcal{P}} \frac{\Pr[S|\mathsf{p}] \Pr[\mathsf{p}]}{\Pr[S]} \\ &= \operatorname*{argmax}_{\mathsf{p} \in \mathcal{P}} \Pr[S|\mathsf{p}] \Pr[\mathsf{p}]. \end{aligned}$$

  - note: for a uniform prior, ML = MAP.

This Lecture

- Notions of information theory.
- Introduction to density estimation.
- Maxent models.
- Conditional Maxent models.

Density Estimation + Features

- Training data: sample S of size m drawn i.i.d. from set $\mathcal{X}$ according to some distribution $\mathcal{D}$,

$$S = (x_1, \ldots, x_m).$$

- Features: associated to elements of $\mathcal{X}$,

$$\begin{aligned} \mathbf{\Phi} \colon \mathcal{X} &\to \mathbb{R}^N \\ x &\mapsto \mathbf{\Phi}(x) = \begin{bmatrix} \Phi_1(x) \\ \vdots \\ \Phi_N(x) \end{bmatrix}. \end{aligned}$$

- Problem: find distribution p out of hypothesis set $\mathcal{P}$ that best estimates $\mathcal{D}$.
  - for simplicity, in what follows, $\mathcal{X}$ is assumed to be finite.

Features

- Feature functions Φ_j assumed to be in H and $\|\mathbf{\Phi}\|_\infty \leq \Lambda$.
- Examples of H:
  - family of threshold functions $\{\mathbf{x} \mapsto 1_{x_i \leq \theta} \colon \mathbf{x} \in \mathbb{R}^N, \theta \in \mathbb{R}\}$ defined over N variables.
  - functions defined via decision trees with larger depths.
  - k-degree monomials of the original features.
  - zero-one features (often used in NLP, e.g., presence/ absence of a word or POS tag).

Maximum Entropy Principle

(E. T. Jaynes, 1957, 1983)

- Idea: empirical feature vector average close to expectation. For any $\delta > 0$, with probability at least $1 - \delta$

$$\left\| \underset{x\sim\mathcal{D}}{\mathrm{E}}[\mathbf{\Phi}(x)] - \underset{x\sim\widehat{\mathcal{D}}}{\mathrm{E}}[\mathbf{\Phi}(x)] \right\|_\infty \leq 2\mathfrak{R}_m(H) + \Lambda\sqrt{\frac{\log\frac{2}{\delta}}{2m}},$$

- Maxent principle: find distribution p that is closest to a prior distribution p_0 (typically uniform distribution) while verifying $\left\| \mathrm{E}_{x\sim\mathsf{p}}[\mathbf{\Phi}(x)] - \mathrm{E}_{x\sim\widehat{\mathcal{D}}}[\mathbf{\Phi}(x)] \right\|_\infty \leq \beta$.
- Closeness is measured using relative entropy.
  - note: no set $\mathcal{P}$ needed to be specified.

Maxent Formulation

- Optimization problem:

$$\min_{\mathsf{p}\in\Delta} D(\mathsf{p} \parallel \mathsf{p}_0)$$

$$\text{subject to: } \left\| \mathop{\mathrm{E}}_{x\sim\mathsf{p}}[\mathbf{\Phi}(x)] - \mathop{\mathrm{E}}_{x\sim S}[\mathbf{\Phi}(x)] \right\|_\infty \leq \beta.$$

  - convex optimization problem, unique solution.
  - $\beta = 0$: standard Maxent (or unregularized Maxent).
  - $\beta > 0$: regularized Maxent.

Relation with Entropy

- Relationship with entropy: for a uniform prior p_0,

$$\begin{aligned} D(\mathsf{p} \parallel \mathsf{p}_0) &= \sum_{x \in \mathcal{X}} \mathsf{p}(x) \log \frac{\mathsf{p}(x)}{\mathsf{p}_0(x)} \\ &= -\sum_{x \in \mathcal{X}} \mathsf{p}(x) \log \mathsf{p}_0(x) + \sum_{x \in \mathcal{X}} \mathsf{p}(x) \log \mathsf{p}(x) \\ &= \log |\mathcal{X}| - H(\mathsf{p}). \end{aligned}$$

Maxent Problem

- Optimization: convex optimization problem.

$$\begin{aligned} \min_{\mathsf{p}} \quad & \sum_{x \in \mathcal{X}} \mathsf{p}(x) \log \mathsf{p}(x) \\ \text{subject to: } & \mathsf{p}(x) \geq 0, \forall x \in \mathcal{X} \\ & \sum_{x \in \mathcal{X}} \mathsf{p}(x) = 1 \\ & \left| \sum_{x \in \mathcal{X}} \mathsf{p}(x) \Phi_j(x) - \frac{1}{m} \sum_{i=1}^{m} \Phi_j(x_i) \right| \leq \beta, \forall j \in [1, N]. \end{aligned}$$

Gibbs Distributions

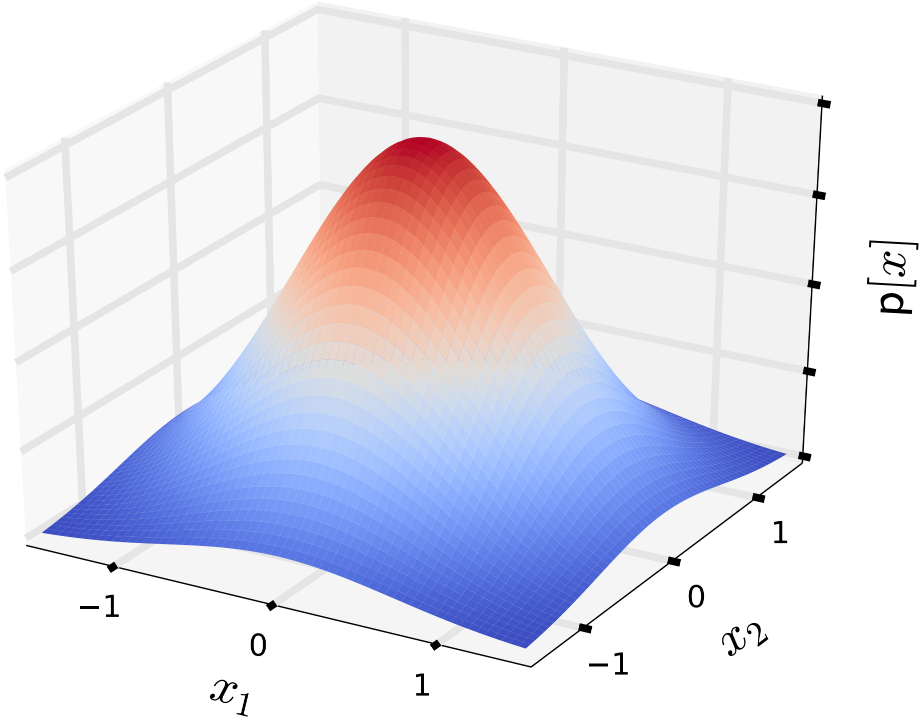
- Gibbs distributions: set $\mathcal{Q}$ of distributions $p_{\mathbf{w}}$ with $\mathbf{w} \in \mathbb{R}^N$,

$$\mathsf{p}_{\mathbf{w}}[x] = \frac{\mathsf{p}_0[x] \exp\big(\mathbf{w} \cdot \boldsymbol{\Phi}(x)\big)}{Z} = \frac{\mathsf{p}_0[x] \exp\big(\sum_{j=1}^N w_j \Phi_j(x)\big)}{Z},$$

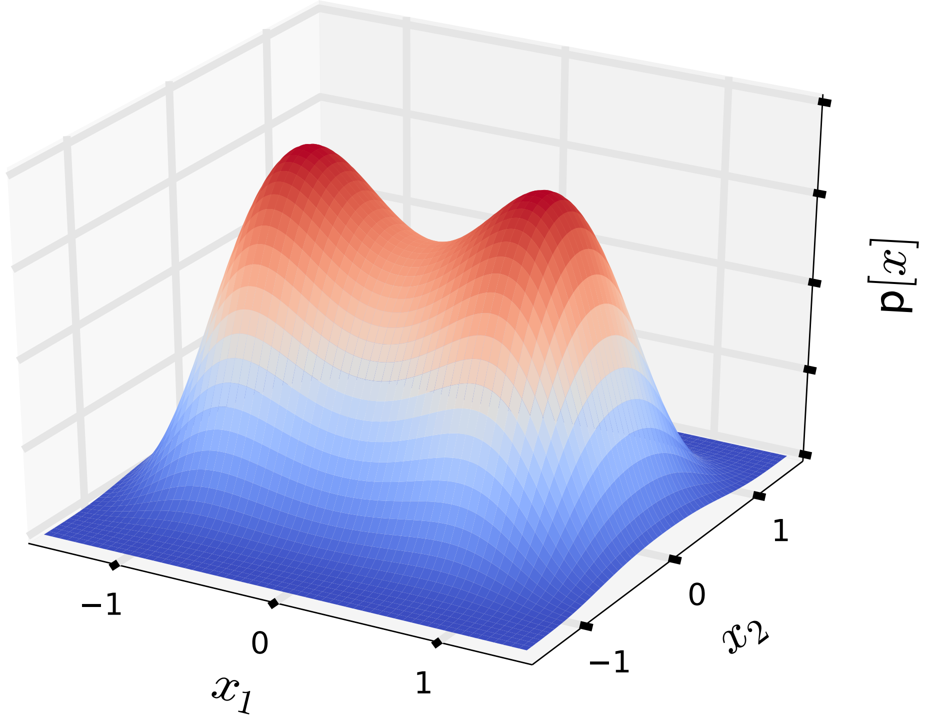
with $Z = \sum_x \mathsf{p}_0[x] \exp\big(\mathbf{w} \cdot \boldsymbol{\Phi}(x)\big)$.

- Rich family:
  - for linear and quadratic features: includes Gaussians and other distributions with non-PSD quadratic forms in exponents.
  - for higher-degree polynomials of raw features: more complex multi-modal distributions.

Examples



$$\mathsf{p}[(x_1, x_2)] = \frac{e^{-(x_1^2+x_2^2)}}{Z}.$$



$$\mathsf{p}[(x_1, x_2)] = \frac{e^{-(x_1^4+x_2^4)+x_1^2-x_2^2}}{Z}.$$

Dual Problems

- Regularized Maxent problem:

$$\min_{\mathsf{p}} F(\mathsf{p}) = \overline{D}(\mathsf{p} \parallel \mathsf{p}_0) + I_C(\operatorname*{E}_{\mathsf{p}}[\mathbf{\Phi}]),$$

$$\text{with } \begin{cases} \overline{D}(\mathsf{p} \parallel \mathsf{p}_0) = D(\mathsf{p} \parallel \mathsf{p}_0) \text{ if } \mathsf{p} \in \Delta\text{, } +\infty \text{ otherwise;} \\ C = \left\{\mathbf{u} \colon \|\mathbf{u} - \operatorname*{E}_{S}[\mathbf{\Phi}]\,\|_\infty \leq \beta\right\}; \\ I_C(x) = 0 \text{ if } x \in C\text{, } I_C(x) = +\infty \text{ otherwise.} \end{cases}$$

- Regularized Maximum Likelihood problem with Gibbs distributions:

$$\sup_{\mathbf{w}} G(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^{m} \log \left[\frac{\mathsf{p}_{\mathbf{w}}[x_i]}{\mathsf{p}_0[x_i]}\right] - \beta \|\mathbf{w}\|_1.$$

Duality Theorem

(Della Pietra et al., 1997; Dudík et al., 2007; Cortes et al., 2015)

- Theorem: the regularized Maxent and ML with Gibbs distributions problems are equivalent,

$$\sup_{\mathbf{w} \in \mathbb{R}^N} G(\mathbf{w}) = \min_{\mathsf{p}} F(\mathsf{p}).$$

  - furthemore, let $\mathsf{p}^* = \operatorname*{argmin}_{\mathsf{p}} F(\mathsf{p})$, then, for any $\epsilon > 0$,

$$\left(\left| G(\mathbf{w}) - \sup_{\mathbf{w} \in \mathbb{R}^N} G(\mathbf{w}) \right| < \epsilon \right) \Rightarrow \Big(D(\mathsf{p}^* \parallel \mathsf{p}_{\mathbf{w}}) \leq \epsilon \Big).$$

Notes

- Maxent formulation:
  - no explicit restriction to a family of distributions $\mathcal{P}$.
  - but solution coincides with regularized ML with a specific family $\mathcal{P}$!
  - more general Bregman divergence-based formulation.

L_1-Regularized Maxent

(Kazama and Tsujii, 2003)

- Optimization problem:

$$\inf_{\mathbf{w}\in\mathbb{R}^N} \beta\|\mathbf{w}\|_1 - \frac{1}{m}\sum_{i=1}^{m} \log \mathsf{p}_{\mathbf{w}}[x_i].$$

$$\text{where } \mathsf{p}_{\mathbf{w}}[x] = \frac{1}{Z}\exp\big(\mathbf{w}\cdot\mathbf{\Phi}(x)\big).$$

- Bayesian interpretation: equivalent to MAP with Laplacian prior $q_{\text{prior}}(\mathbf{w})$ (Williams, 1994),

$$\max_{\mathbf{w}} \log\Big(\prod_{i=1}^{m} \mathsf{p}_{\mathbf{w}}[x_i]\, \mathsf{q}_{\text{prior}}(\mathbf{w})\Big)$$

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

Generalization Guarantee

(Dudík et al., 2007)

- Notation: $\mathcal{L}_{\mathcal{D}}(\mathbf{w}) = \mathop{\mathrm{E}}_{x \sim \mathcal{D}}[-\log \mathsf{p}_{\mathbf{w}}[x]]$, $\mathcal{L}_S(\mathbf{w}) = \mathop{\mathrm{E}}_{x \sim S}[-\log \mathsf{p}_{\mathbf{w}}[x]]$.
- Theorem: Fix $\delta > 0$. Let $\widehat{\mathbf{w}}$ be the solution of the L1-reg. Maxent problem for $\beta = 2\mathfrak{R}_m(H) + \Lambda\sqrt{\log(\frac{2}{\delta})/2m}$. Then, with probability at least $1 - \delta$,

$$\mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}}) \leq \inf_{\mathbf{w}} \mathcal{L}_{\mathcal{D}}(\mathbf{w}) + 2\|\mathbf{w}\|_1 \left[2\mathfrak{R}_m(H) + \Lambda\sqrt{\frac{\log \frac{2}{\delta}}{2m}}\right].$$

Proof

- By Hölder's inequality and the concentration bound for average feature vectors,

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}}) - \mathcal{L}_S(\widehat{\mathbf{w}}) &= \widehat{\mathbf{w}} \cdot \big[\mathop{\mathrm{E}}_{S}[\mathbf{\Phi}] - \mathop{\mathrm{E}}_{\mathcal{D}}[\mathbf{\Phi}]\big] \\ &\leq \|\widehat{\mathbf{w}}\|_1 \, \|\mathop{\mathrm{E}}_{S}[\mathbf{\Phi}] - \mathop{\mathrm{E}}_{\mathcal{D}}[\mathbf{\Phi}]\|_\infty \leq \beta \|\widehat{\mathbf{w}}\|_1.\end{aligned}$$

- Since $\widehat{\mathbf{w}}$ is a minimizer,

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}}) - \mathcal{L}_{\mathcal{D}}(\mathbf{w}) &= \mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}}) - \mathcal{L}_S(\widehat{\mathbf{w}}) + \mathcal{L}_S(\widehat{\mathbf{w}}) - \mathcal{L}_{\mathcal{D}}(\mathbf{w}) \\ &\leq \beta \|\widehat{\mathbf{w}}\|_1 + \mathcal{L}_S(\widehat{\mathbf{w}}) - \mathcal{L}_{\mathcal{D}}(\mathbf{w}) \\ &\leq \beta \|\mathbf{w}\|_1 + \mathcal{L}_S(\mathbf{w}) - \mathcal{L}_{\mathcal{D}}(\mathbf{w}) \leq 2\beta \|\mathbf{w}\|_1. \\ &\qquad (\widehat{\mathbf{w}} \text{ minimizer of } \beta\|\mathbf{w}\|_1 + \mathcal{L}_S(\mathbf{w}))\end{aligned}$$

L_2-Regularized Maxent

(Chen and Rosenfeld, 2000; Lebanon and Lafferty, 2001)

- Different relaxations:
  - L_1 constraints:

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

  - L_2 constraints:

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

L_2-Regularized Maxent

- Optimization problem:

$$\inf_{\mathbf{w} \in \mathbb{R}^N} \beta \|\mathbf{w}\|_2^2 - \frac{1}{m} \sum_{i=1}^{m} \log \mathsf{p}_{\mathbf{w}}[x_i].$$

$$\text{where } \mathsf{p}_{\mathbf{w}}[x] = \frac{1}{Z} \exp\big(\mathbf{w} \cdot \mathbf{\Phi}(x)\big).$$

- Bayesian interpretation: equivalent to MAP with Gaussian prior $q_{\text{prior}}(\mathbf{w})$ (Goodman, 2004),

$$\max_{\mathbf{w}} \ \log\Big(\prod_{i=1}^{m} \mathsf{p}_{\mathbf{w}}[x_i]\, \mathsf{q}_{\text{prior}}(\mathbf{w})\Big)$$

$$\text{with } \mathsf{q}_{\text{prior}}(\mathbf{w}) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{w_j^2}{2\sigma^2}}.$$

This Lecture

- Notions of information theory.
- Introduction to density estimation.
- Maxent models.
- Conditional Maxent models.

Conditional Maxent Models

- Maxent models for conditional probabilities:
  - conditional probability modeling each class.
  - use in multi-class classification.
  - can use different features for each class.
  - a.k.a. multinomial logistic regression.
  - logistic regression: special case of two classes.

Problem

- Data: sample drawn i.i.d. according to some distribution D,

$$S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m.$$

  - $\mathcal{Y} = \{1, \ldots, k\}$, or $\mathcal{Y} = \{0, 1\}^k$ in multi-label case.

- Features: mapping $\mathbf{\Phi} \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^N$.

- Problem: find accurate conditional probability models $\Pr[\cdot \mid x]$, $x \in \mathcal{X}$, based on $\mathbf{\Phi}$.

Conditional Maxent Principle

(Berger et al., 1996; Cortes et al., 2015)

- Idea: empirical feature vector average close to expectation. For any $\delta > 0$, with probability at least $1 - \delta$,

$$\left\| \underset{\substack{x\sim\widehat{\mathsf{p}}\\ y\sim\mathcal{D}[\cdot|x]}}{\mathrm{E}}[\boldsymbol{\Phi}(x,y)] - \underset{\substack{x\sim\widehat{\mathsf{p}}\\ y\sim\widehat{\mathsf{p}}[\cdot|x]}}{\mathrm{E}}[\boldsymbol{\Phi}(x,y)] \right\|_\infty \leq 2\mathfrak{R}_m(H) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

- Maxent principle: find conditional distributions $\mathsf{p}[\cdot|x]$ that are closest to priors $\mathsf{p}_0[\cdot|x]$ (typically uniform distributions) while verifying $\left\| \mathrm{E}_{\substack{x\sim\widehat{\mathsf{p}}\\ y\sim\mathsf{p}[\cdot|x]}}[\boldsymbol{\Phi}(x,y)] - \mathrm{E}_{\substack{x\sim\widehat{\mathsf{p}}\\ y\sim\widehat{\mathsf{p}}[\cdot|x]}}[\boldsymbol{\Phi}(x,y)] \right\|_\infty \leq \beta$.

- Closeness is measured using conditional relative entropy based on $\widehat{\mathsf{p}}$.

Cond. Maxent Formulation

(Berger et al., 1996; Cortes et al., 2015)

- Optimization problem: find distribution p solution of

$$\min_{\mathsf{p}[\cdot|x] \in \Delta} \sum_{x \in \mathcal{X}} \widehat{\mathsf{p}}[x] \, D\big(\mathsf{p}[\cdot|x] \parallel \mathsf{p}_0[\cdot|x]\big)$$

$$\text{s.t.} \; \left\| \mathop{\mathrm{E}}_{x \sim \widehat{\mathsf{p}}} \left[\mathop{\mathrm{E}}_{y \sim \mathsf{p}[\cdot|x]} [\mathbf{\Phi}(x, y)] \right] - \mathop{\mathrm{E}}_{(x,y) \sim S} [\mathbf{\Phi}(x, y)] \right\|_\infty \leq \beta.$$

  - convex optimization problem, unique solution.
  - $\beta = 0$: unregularized conditional Maxent.
  - $\beta > 0$: regularized conditional Maxent.

Dual Problems

- Regularized conditional Maxent problem:

$$\widetilde{F}(\mathsf{p}) = \mathop{\mathrm{E}}_{x\sim\widehat{p}}\left[\overline{D}\big(\mathsf{p}[\cdot|x] \parallel \mathsf{p}_0[\cdot|x]\big) + I_\Delta\big(\mathsf{p}[\cdot|x]\big)\right] + I_C\Big(\mathop{\mathrm{E}}_{\substack{x\sim\widehat{\mathsf{p}}\\ y\sim\mathsf{p}[\cdot|x]}}[\mathbf{\Phi}]\Big).$$

- Regularized Maximum Likelihood problem with conditional Gibbs distributions:

$$\widetilde{G}(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m}\log\left[\frac{\mathsf{p}_{\mathbf{w}}[y_i|x_i]}{\mathsf{p}_0[y_i|x_i]}\right] - \beta\|\mathbf{w}\|_1\,,$$

where $\forall(x,y)\in\mathcal{X}\times\mathcal{Y}$,

$$\mathsf{p}_{\mathbf{w}}[y|x] = \frac{\mathsf{p}_0[y|x]\exp\big(\mathbf{w}\cdot\mathbf{\Phi}(x,y)\big)}{Z(x)}$$

$$Z(x) = \sum_{y\in\mathcal{Y}}\mathsf{p}_0[y|x]\exp(\mathbf{w}\cdot\mathbf{\Phi}(x,y)).$$

Duality Theorem

(Cortes et al., 2015)

- Theorem: the regularized conditional Maxent and ML with conditional Gibbs distributions problems are equivalent,

$$\sup_{\mathbf{w}\in\mathbb{R}^N} \widetilde{G}(\mathbf{w}) = \min_{\mathsf{p}} \widetilde{F}(\mathsf{p}).$$

  - furthemore, let $\mathsf{p}^* = \operatorname*{argmin}_{\mathsf{p}} \widetilde{F}(\mathsf{p})$, then, for any $\epsilon > 0$,

$$\Big(|\widetilde{G}(\mathbf{w}) - \sup_{\mathbf{w}\in\mathbb{R}^N} \widetilde{G}(\mathbf{w})| < \epsilon\Big) \Rightarrow \mathop{\mathrm{E}}_{x\sim\widehat{p}}\Big[D\big(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_{\mathbf{w}}[\cdot|x]\big)\Big] \leq \epsilon.$$

Regularized Cond. Maxent

(Berger et al., 1996; Cortes et al., 2015)

- Optimization problem: convex optimizations, regularization parameter $\lambda \geq 0$.

$$\min_{\mathbf{w} \in \mathbb{R}^N} \lambda \|\mathbf{w}\|_1 - \frac{1}{m} \sum_{i=1}^{m} \log \mathsf{p}_{\mathbf{w}}[y_i | x_i]$$

$$\text{or} \quad \min_{\mathbf{w} \in \mathbb{R}^N} \lambda \|\mathbf{w}\|_2^2 - \frac{1}{m} \sum_{i=1}^{m} \log \mathsf{p}_{\mathbf{w}}[y_i | x_i],$$

where $\forall (x, y) \in \mathcal{X} \times \mathcal{Y},$

$$\mathsf{p}_{\mathbf{w}}[y|x] = \frac{\exp\big(\mathbf{w} \cdot \mathbf{\Phi}(x, y)\big)}{Z(x)}$$

$$Z(x) = \sum_{y \in \mathcal{Y}} \exp\big(\mathbf{w} \cdot \mathbf{\Phi}(x, y)\big).$$

More Explicit Forms

- Optimization problem: multinomial logistic loss.

$$\min_{\mathbf{w}\in\mathbb{R}^N} \begin{cases} \lambda\|\mathbf{w}\|_1 \\ \lambda\|\mathbf{w}\|_2^2 \end{cases} + \frac{1}{m}\sum_{i=1}^{m} \log\left[\sum_{y\in\mathcal{Y}} \exp\Big(\mathbf{w}\cdot\mathbf{\Phi}(x_i, y) - \mathbf{w}\cdot\mathbf{\Phi}(x_i, y_i)\Big)\right].$$

$$\min_{\mathbf{w}\in\mathbb{R}^N} \begin{cases} \lambda\|\mathbf{w}\|_1 \\ \lambda\|\mathbf{w}\|_2^2 \end{cases} - \mathbf{w}\cdot\frac{1}{m}\sum_{i=1}^{m}\mathbf{\Phi}(x_i, y_i) + \frac{1}{m}\sum_{i=1}^{m}\log\left[\sum_{y\in\mathcal{Y}} e^{\mathbf{w}\cdot\mathbf{\Phi}(x_i, y)}\right].$$

Related Problem

- Optimization problem: log-sum-exp replaced by max.

$$\min_{\mathbf{w}\in\mathbb{R}^N} \begin{cases} \lambda\|\mathbf{w}\|_1 \\ \lambda\|\mathbf{w}\|_2^2 \end{cases} + \frac{1}{m}\sum_{i=1}^{m} \underbrace{\max_{y\in\mathcal{Y}}\Big(\mathbf{w}\cdot\mathbf{\Phi}(x_i,y) - \mathbf{w}\cdot\mathbf{\Phi}(x_i,y_i)\Big)}_{-\rho_{\mathbf{w}}(x_i,y_i)}.$$

Common Feature Choice

- Multi-class features:

$$\mathbf{\Phi}(x,y) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \mathbf{\Gamma}(x) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_{y-1} \\ \mathbf{w}_y \\ \mathbf{w}_{y+1} \\ \vdots \\ \mathbf{w}_{|\mathcal{Y}|} \end{bmatrix} \quad \longrightarrow \quad \mathbf{w} \cdot \mathbf{\Phi}(x,y) = \mathbf{w}_y \cdot \mathbf{\Gamma}(x).$$

- L2-regularized cond. maxent optimization:

$$\min_{\mathbf{w} \in \mathbb{R}^N} \lambda \sum_{y \in \mathcal{Y}} \|\mathbf{w}_y\|_2^2 + \frac{1}{m} \sum_{i=1}^{m} \log \left[\sum_{y \in \mathcal{Y}} \exp \Big(\mathbf{w}_y \cdot \mathbf{\Gamma}(x_i) - \mathbf{w}_{y_i} \cdot \mathbf{\Gamma}(x_i) \Big) \right].$$

Prediction

- Prediction with $\mathsf{p}_{\mathbf{w}}[y|x] = \frac{\exp(\mathbf{w}\cdot\mathbf{\Phi}(x,y))}{Z(x)}$:

$$\widehat{y}(x) = \text{argmax}_{y\in\mathcal{Y}}\ \mathsf{p}_{\mathbf{w}}[y|x] = \text{argmax}_{y\in\mathcal{Y}}\ \mathbf{w}\cdot\mathbf{\Phi}(x,y).$$

Binary Classification

- Simpler expression:

$$\begin{aligned}
&\sum_{y\in\mathcal{Y}} \exp\Big(\mathbf{w}\cdot\mathbf{\Phi}(x_i, y) - \mathbf{w}\cdot\mathbf{\Phi}(x_i, y_i)\Big)\\
&= e^{\mathbf{w}\cdot\mathbf{\Phi}(x_i,+1)-\mathbf{w}\cdot\mathbf{\Phi}(x_i,y_i)} + e^{\mathbf{w}\cdot\mathbf{\Phi}(x_i,-1)-\mathbf{w}\cdot\mathbf{\Phi}(x_i,y_i)}\\
&= 1 + e^{-y_i\mathbf{w}\cdot[\mathbf{\Phi}(x_i,+1)-\mathbf{\Phi}(x_i,-1)]}\\
&= 1 + e^{-y_i\mathbf{w}\cdot\mathbf{\Psi}(x_i)},
\end{aligned}$$

with $\mathbf{\Psi}(x) = \mathbf{\Phi}(x, +1) - \mathbf{\Phi}(x, -1)$.

Logistic Regression

(Berkson, 1944)

- Binary case of conditional Maxent.
- Optimization problem: regularized logistic loss.

$$\min_{\mathbf{w}\in\mathbb{R}^N} \begin{cases} \lambda\|\mathbf{w}\|_1 \\ \lambda\|\mathbf{w}\|_2^2 \end{cases} + \frac{1}{m}\sum_{i=1}^{m} \log\left[1 + e^{-y_i \mathbf{w}\cdot\mathbf{\Psi}(x_i)}\right].$$

  - convex optimization.
  - variety of solutions: SGD, coordinate descent, etc.
  - coordinate descent: similar to AdaBoost with logistic loss $\phi(-u) = \log_2(1 + e^{-u}) \geq 1_{u\leq 0}$ instead of exponential loss.

Generalization Bound

- Theorem: assume that $\pm\Phi_j \in H$ for all $j \in [1, N]$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of a sample S of size m, for all $f\colon x \mapsto \mathbf{w} \cdot \Phi(x)$,

$$R(f) \leq \frac{1}{m}\sum_{i=1}^{m} \log_{u_0}\Big(1 + e^{-y_i \mathbf{w}\cdot\Phi(x_i)}\Big) + 4\|\mathbf{w}\|_1 \mathfrak{R}_m(H)$$
$$+ \sqrt{\frac{\log\log_2 2\|\mathbf{w}\|_1}{m}} + \sqrt{\frac{\log\frac{2}{\delta}}{m}},$$

where $u_0 = 1 + \frac{1}{e}$.

Proof

- Proof: by the learning bound for convex ensembles holding uniformly for all ρ, with probability at least $1-\delta$, for all f and $\rho > 0$,

$$R(f) \leq \frac{1}{m} \sum_{i=1}^{m} 1_{\frac{y_i \mathbf{w} \cdot \Phi(x_i)}{\rho \|\mathbf{w}\|_1} - 1 \leq 0} + \frac{4}{\rho} \mathfrak{R}_m(H) + \sqrt{\frac{\log \log_2 \frac{2}{\rho}}{m}} + \sqrt{\frac{\log \frac{2}{\delta}}{m}}.$$

- Choosing $\rho = \frac{1}{\|\mathbf{w}\|_1}$ and using $1_{u \leq 1} \leq \log_{u_0}(1 + e^{-u})$ yields immediately the learning bound of the theorem.

Logistic Regression

(Berkson, 1944)

- Logistic model:

$$\Pr[y=+1 \mid x] = \frac{e^{\mathbf{w}\cdot\mathbf{\Phi}(x,+1)}}{Z(x)},$$

where $Z(x) = e^{\mathbf{w}\cdot\mathbf{\Phi}(x,+1)} + e^{\mathbf{w}\cdot\mathbf{\Phi}(x,-1)}$

- Properties:
  - linear decision rule, sign of log-odds ratio:

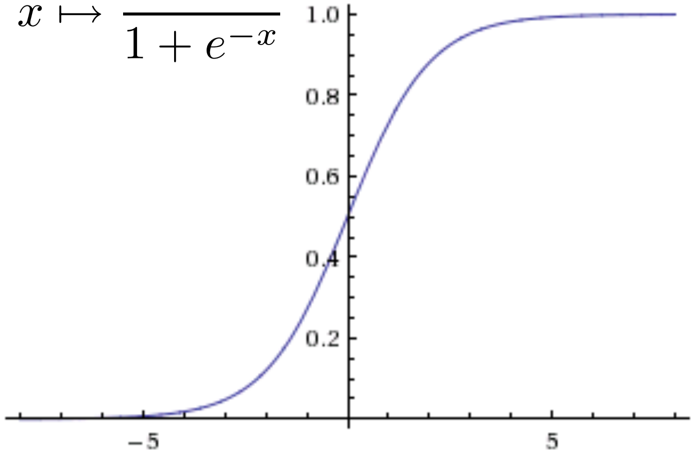
$$\log \frac{\Pr[y=+1 \mid x]}{\Pr[y=-1 \mid x]} = \mathbf{w} \cdot \big(\mathbf{\Phi}(x,+1) - \mathbf{\Phi}(x,-1)\big) = \mathbf{w} \cdot \mathbf{\Psi}(x).$$

  - logistic form:

$$\Pr[y=+1 \mid x] = \frac{1}{1 + e^{-\mathbf{w}\cdot[\mathbf{\Phi}(x,+1)-\mathbf{\Phi}(x,-1)]}} = \frac{1}{1 + e^{-\mathbf{w}\cdot\mathbf{\Psi}(x)}}.$$

Logistic/Sigmoid Function

$$f\colon x \mapsto \frac{1}{1+e^{-x}}$$



$$\Pr[y = +1 \mid x] = f\big(\mathbf{w} \cdot \mathbf{\Psi}(x)\big).$$

Applications

- Natural language processing (Berger et al., 1996; Rosenfeld, 1996; Pietra et al., 1997; Malouf, 2002; Manning and Klein, 2003; Mann et al., 2009; Ratnaparkhi, 2010).
- Species habitat modeling (Phillips et al., 2004, 2006; Dudík et al., 2007; Elith et al, 2011).
- Computer vision (Jeon and Manmatha, 2004).

Extensions

- Extensive theoretical study of alternative regularizations: (Dudík et al., 2007) (see also (Altun and Smola, 2006) though some proofs unclear).
- Maxent models with other Bregman divergences (see for example (Altun and Smola, 2006)).
- Structural Maxent models (Cortes et al., 2015):
    - extension to the case of multiple feature families.
    - empirically outperform Maxent and L1-Maxent.
    - conditional structural Maxent: coincide with deep boosting using the logistic loss.

Conclusion

- Logistic regression/maxent models:
  - theoretical foundation.
  - natural solution when probabilites are required.
  - widely used for density estimation/classification.
  - often very effective in practice.
  - distributed optimization solutions.
  - no natural non-linear L1-version (use of kernels).
  - connections with boosting.
  - connections with neural networks.

References

- Yasemin Altun, Alexander J. Smola. Unifying Divergence Minimization and Statistical Inference Via Convex Duality. COLT 2006: 139-153

- 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.

- Lev M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Computational Mathematics and Mathematical Physics, 7:200–217, 1967.

- Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, and Umar Syed. Structural Maxent. In Proceedings of ICML, 2015.

References

- 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.

- Dudík, Miroslav, Phillips, Steven J., and Schapire, Robert E. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. JMLR, 8, 2007.

References

- 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.

- Solomon Kullback and Richard A. Leibler. On information and sufficiency. Ann. Math. Statist., 22(1):79–86, 1951.

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

References

- Alfréd Rényi. On measures of entropy and information. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, pages 547–561. University of California Press, 1961.

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

- Claude E. Shannon. A mathematical theory of communication. Bell System Technical Journal, 27:379423, 1948.