Entropy

Emprirical Distribution:

$$\widehat{p}_x = \frac{\#\{i : X_i = x\}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i = x\}$$

Stirling's Approximation:

$$k! \approx k^k e^{-k}\sqrt{2\pi k}$$

Shannon Entropy: For p a distribution,

$$H(p) = -\sum_x p(x)\log p(x)$$

with

- $H(X,Y) = H(X) + H(Y)$ if X and Y are independent

Maximum Entropy Principle:

$$p(x) = \frac{1}{Z}\exp\left(\sum_{i=1}^{k}\lambda_i T_i(x)\right)$$

for normalizing constant Z and parameters $\lambda_{i=1:k}$ is the distribution that maximizes $H(p)$ subject to

$$\begin{cases} \sum p_x \mathcal{E}_1(x) = \theta_1 \\ \sum p_x \mathcal{E}_2(x) = \theta_2 \\ \vdots \\ \sum p_x \mathcal{E}_k(x) = \theta_k \\ \sum p_x = 1 \end{cases}$$

Large Deviation Principle: Let p be a distribution on $\{1,\dots,s\}$ and $\mathcal{E} : \{1,\dots,s\} \to \mathbb{R}$. If $X_1,\dots,X_n \overset{extiid}{\sim} p$ satisfy

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{E}(X_i) = \theta$$

then

$$\mathbb{E}_{\widehat{p}}\mathcal{E}(X) = \sum_{x=1}^{s}\widehat{p}_x\mathcal{E}(x) = \theta$$

KL Divergence: For p, q two distributions,

$$D(q \,\|\, p) = \sum_{x=1}^{s} q_x \log\frac{q_x}{p_x}$$

satisfies

1. $D(q \,\|\, p) \geq 0$
2. $D(q \,\|\, p) = 0 \iff q = p$

Convexity: f is convex if $\forall\lambda \in [0,1]$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$

equivalently, if $f''(x) \geq 0$

Jensen's Inequality: For $g : \mathbb{R} \to \mathbb{R}$ convex, $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$

Sanov's Theorem: Let B be an open subset of the set of function on $\{1,\dots,s\}$. Then

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}(\widehat{p} \in B) = -\inf_{q\in B} D(q \,\|\, p)$$

where $\widehat{q}$ is the empirical distribution of $X_1,\dots,X_n$.

Further, if $p^* = \arg\min_{q\in B} D(q \,\|\, p)$ is unique, then

$$\lim_{n\to\infty}\mathbb{P}(||\widehat{p} - p^*|| > \varepsilon \mid \widehat{p} \in B) = 0 \qquad \forall\varepsilon > 0$$

for any metric $||\cdot||$ on the space of distributions.

Exponential Family: Distributions of the form

$$p(x;\lambda) = \frac{1}{Z(\lambda)}\exp\left(\sum_{i=1}^{k}\lambda_i T_i(x)\right)$$

comprise an exponential family with sufficient statistics $T_i(x)$ and natural parameters λ_i satisfying

1. $\frac{\partial}{\partial\lambda_k}\log Z_\lambda = \mathbb{E}_\lambda[T_k(x)]$
2. $\frac{\partial^2}{\partial\lambda_k\,\partial\lambda_j}\log Z_\lambda = \mathrm{Cov}_\lambda(T_k(x), T_j(x))$
3. $\log Z_\lambda$ is convex in λ and strictly convex unless the conditions $\{E_{p^*}[T_k(x)] = \theta_k\}_{k=1}^{c}$ are redundant
4. $\log Z_\lambda - \sum_{k=1}^{c}\lambda_k\theta_k$ is strictly convex and is minimized when $\mathbb{E}_\lambda[T_k(x)] = \theta_k$

Source Coding

Prefix code: A *code* $C : \{1,\dots,t\} \to \{0,1\}^*$ is a *prefix code* if $C(x)$ is not a prefix of $C(y)$ for any $x \neq y$.

Kraft-McMillan:

1. For all prefix codes C,
$$\sum_{x=1}^{t} 2^{-|C(x)|} \leq 1$$
2. For any code lengths $\ell_1,\dots,\ell_t$ satisfying
$$\sum_{x=1}^{t} 2^{-\ell_x} \leq 1$$
there exists a prefix code C such that $|C(x)| = \ell_x$ for all $x = 1 : t$.

Theorem: Let $X \sim p$. For the optimal $C^* = \arg\min_{C \text{ prefix}} \mathbb{E}_p\,|C(x)|$,

$$H(p) \leq \mathbb{E}_p\,|C^*(X)| \leq H(p) + 1$$

Block Coding: Further, for n fixed,

$$H(p) \leq \frac{1}{n}\mathbb{E}_p\,|C_n^*(X_{1:n})| \leq H(p) + \frac{1}{n}$$

so by coding large enough blocks, we can get arbitrarily close to $H(p)$ bits/symbol.

Statistical Learning

Unbiased Estimator: Suppose $\widehat{\theta} = \widehat{\theta}(X_1,\dots,X_n)$ is an estimator of θ. We say $\widehat{\theta}$ is unbiased if $\mathbb{E}[\widehat{\theta}] = \theta$

Consistency: $\widehat{\theta}_n$ is consistent if $\widehat{\theta}_n \to \theta$ in some sense.

- $\widehat{\theta}_n \xrightarrow{a.s.} \theta$ if $\mathbb{P}(\lim_{n\to\infty}\widehat{\theta}_n = \theta) = 1$
- $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta$ if $\forall\varepsilon > 0$, $\mathbb{P}\left(\left|\widehat{\theta}_n - \theta\right| > \varepsilon\right) \to 0$ as $n \to \infty$
- $\widehat{\theta}_n \xrightarrow{L^2} \theta$ if $\mathbb{E}\left[\left|\widehat{\theta}_n - \theta\right|^2\right] \to 0$ as $n \to \infty$

Mean Square Error (MSE): $\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}\left|\widehat{\theta}_n - \theta\right|^2 = \mathrm{Var}(\widehat{\theta}) + (\mathbb{E}[\widehat{\theta}_n] - \theta)^2$

Kensity Density Estimation: For a function k satisfying $k \geq 0$, $E[k] = 0$, $\mathrm{Var}[k] = 1$, we approximate a discrete density f by the continuous density

$$\begin{aligned}\widehat{f}_{n,w}(x, X_1,\dots,X_n) &= \frac{1}{n}\sum_{i=1}^{n} k_w(x - X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{k(x/w)}{w}\end{aligned}$$

since for $\mathrm{MSE} = \mathrm{bias}^2 + \mathrm{variance}$, we have

- bias $\searrow$ and variance $\nearrow$ as $w \to 0$
- bias $\nearrow$ and variance $\searrow$ as $w \to \infty$

and

$$\mathbb{E}[\widehat{f}_{n,w}(x)] = \int_{\mathbb{R}} f(t)k_w(x-t)\,dt = (f \star k_w)(x)$$

Integrated Square Error (ISE):

$$\mathrm{ISE} = \int_{\mathbb{R}}\left|\widehat{f}_n(x,\ X_1,\dots,X_n) - f(x)\right|^2\,dx$$

Mean Integrated Square Error (MISE):

$$\mathrm{MISE} = \mathbb{E}[\mathrm{ISE}] = \int_{\mathbb{R}}\mathbb{E}\left|\widehat{f}(x,\ X_1,\dots,X_n)\right|^2\,dx$$

Theorem: For f smooth and k a kernel density, as $w \to 0$,

$$\mathrm{MISE}_{n,w} = \alpha w^4 + \frac{\beta}{nw} + \mathrm{error}$$

for α, β constants.

Sylverman's Rule of Thumb: The optimal bandwidth $w^* \propto n^{-1/5}$.

Cross-validation Estimator: With $\widehat{f}_{n-1,w}^{(i)}(X_i) = \widehat{f}_{n-1,w}(x,\ X_1 \dots X_{i-1}, X_{i+1} \dots X_n)$ define

$$I = \frac{1}{n}\sum_{i=1}^{n}\widehat{f}_{n-1,w}^{(i)}(X_i)$$

Theorem (Stone 1984): with w chosen by

$$\arg\min_w\left[\int\widehat{f}_{n,w}(x)^2 - \frac{2}{n}\sum_{i=1}^{n}\widehat{f}_{n-1,w}^{(i)}(X_i)\right]$$

we have

$$\mathrm{ISE}(\widehat{f}_{\widehat{w}_n}, f) \xrightarrow{a.s.} \inf_w \mathrm{ISE}(\widehat{f}_{w,n}, f)$$

though the convergence is very slow in high-dimensional spaces.

Maximum Likelihood Estimation (MLE):

$$\begin{aligned}\widehat{\theta} &= \arg\max_\theta p_\theta(X_1 = x_1,\dots,X_n = x_n) \\ &= \arg\min_\theta D(\widehat{p} \,\|\, p_\theta)\end{aligned}$$

Bayes' Classification Rule:

$$h^*(x) = \arg\max_c \mathbb{P}(Y = C \mid X = x)$$
$$= \arg\max_c \frac{\pi_c f_c(x)}{\mathbb{P}(X = x)}$$

for $\pi_i = \mathbb{P}(Y = i)$, $f_i(x) = \mathbb{P}(X = x \mid Y = i)$ the class-conditional densities.

Neyman-Pearson Classification: Fix $t \in (0, \infty)$. Then

$$h_t(x) = \begin{cases} 2 & \frac{\pi_1 f_1(x)}{\pi_2 f_2(x)} > t \\ 1 & \text{otherwise} \end{cases}$$

Remark: In the case $t = \pi_1/\pi_2$, NP is equivalent to Bayes' classification rule (the optimal classifier).

Theorem: For h any classifier, with $\mathbb{P}(h(X) = 2 \mid Y = 1) \leq \mathbb{P}(h_{NP}(X) = 2 \mid Y = 1)$, we have $\mathbb{P}(h(X) = 2 \mid Y = 2) \leq \mathbb{P}(h_{NP}(X) = 2 \mid Y = 2)$.

That is, NP is the classifier which maximizes the detection rate relative to the false alarm rate.

Naive Bayes': Assume that $f_c(x_1, \ldots, x_d) = \prod_{i=1}^d f_c(x_i)$.

Softmax: Let $r_c(x) = \mathbb{P}(Y = c \mid X = x)$. Then we have linear decision boundaries

$$\log \frac{r_k(x)}{r_1(x)} = \alpha_k + \beta_k x$$

and

$$r_k(x) = \frac{e^{\alpha_k + \beta_k x}}{1 + \sum_{k=2}^s e^{\alpha_k + \beta_k x}}$$

where we find α_k, β_k by MLE.

k-Nearest Neighbors: Let $D_k(x)$ be the closed ball at x with radius $R_k(x)$, the smallest radius that contains k points. Then

$$\widehat{r}_c(x) = \frac{\#\{i : X_i \in D_k(x),\ Y_i = c\}}{k}$$

In this case,

$$\widehat{r}_c(x) \to r_c(x)$$

i.e., the estimator is consistent.

Support Vector Machine: For any collection of data $\{(X_i, Y_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \mathbb{Z}_2$, we can find a transformation $\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$ with $d' \gg d$, such that the d'-dimensional hyperplane $\alpha + \beta x_i$ separates the data.

Our goal then is to find the **maximum margin classifier**

$$h(x) = \operatorname{sign}(\widehat{\alpha} + \widehat{\beta} x)$$

where

$$(\widehat{\alpha}, \widehat{\beta}) = \arg\max_{\alpha,\beta} \min_{i=1:n} \operatorname{dist}(X_i, \{\alpha + \beta x = 0\})$$

for all $i : (\alpha + \beta x_i) Y_i \geq 0$.

Graphical Models

Clique: Let $G = (V, E)$ be a graph. Then $C \subseteq V$ is a *clique* if $\forall i \neq j \in C$, $(i, j) \in E$.

Gibbs Random Field (GRF): $\{X_v\}_{v \in V}$ is a GRF with respect to G if

$$p(x) = \frac{1}{Z} \prod_{c \text{ cliques in } G} \phi_c(x_c)$$

for some $\phi_c : \Omega_c \to [0, \infty)$ clique functions and Z a partition function.

Strictly Positive GRF: If $\phi_c > 0$ for all c, then the GRF is *strictly positive*. Equivalently, $\forall x_1, \ldots, x_M$, $p(x_1, \ldots, x_M) > 0$.

Markov Chain: A Markov chain satisfies

$$p(x_1, \ldots, x_n) = p(x_1) \prod_{i=2}^n p(x_i \mid x_{i-1})$$

Proposition (Independence): Two random variables X and Y on a GRF are independent if there exists no paths between them

---

Remark: Independence does not imply there is no path between X and Y (even on a minimal graph!)

Conditioning Theorem: Let $A \subseteq V(G)$ be a set of nodes. Then $(X_v)_{v \in A}$,conditioned on $X_{V \setminus A}$, is a GRF with respect to the subgraph

$$G|_A = (A, \{(i, j) : i, j \in A,\ i \overset{G}{\sim} j\})$$

Marginalizing Theorem: Let $A \subseteq V(G)$ be a set of nodes. Then $\{X_v\}_{v \in A}$, marginalized over $\{X_{V \setminus A}\}$, is a GRF with respect to the graph $G' = (A, E')$ where

$$u \overset{G'}{\sim} v \iff \begin{cases} u \overset{G}{\sim} v \\ \text{exists path from } u \text{ to } v \text{ in } A^c \end{cases}$$

Markov Random Field (MRF): $(X_v)_{v \in G}$ is a Markov Random Field if

$$\mathbb{P}(X_i = x_i \mid X_{i^c} = x_{i^c}) = \mathbb{P}(X_i = x_i \mid x_{N(i)} = x_{N(i)})$$

where $N(i) = \{j : (i, j) \in E\}$ is the neighborhood of i in G.

Theorem (Hammersley-Clifford): Assume (X_v) is strictly positive. Then X is a GRF iff it is a MRF.

Dynamic Programming: To sample from a GRF, we need to know the partition function Z. We can calculate this by

$$Z = \sum_{x_v} \prod_{c \text{ cliques}} \phi_c(x_c)$$

or by the much faster

1. Sample X_1
2. Sample $X_2 \mid X_1$
3. Sample $X_n \mid X_1, \ldots, X_{n-1}$.

according to the visitation schedule that minimizes $\sum_x |\Omega|^{k_x}$ where $k_x = \#\text{new neighbors} + 1$ and $|\Omega|$ is the size of the state space.

Gibbs Sampling: Gibbs Sampling provides a cost effective alternative to dynamic programming:

1. Randomly initialize $X_1^{(0)}, \ldots, X_n^{(0)}$
2. Sample a vertex $i \sim \pi$ where π is any distribution over V
3. Let $X_i^{(t)} \sim p(x_i^{(t-1)} \mid x_{i^c}^{(t-1)})$ and $X_{i^c}^{(t)} = X_{i^c}^{(t-1)}$
4. Iterate

Proposition: Let $X^{(0)}, \ldots, X^{(N)}$ be a Gibbs sampler and q_t a distribution on $X^(t)$. Then

$$D(q_t \parallel p) \leq D(q_{t-1} \parallel p) \qquad \forall t$$

EM Algorithm: For a general exponential family

$$f(x, y, \lambda) = \frac{1}{Z_\lambda} p(x, y) e^{\sum_{i=1}^k \lambda_i T_i(x,y)}$$

wiht observed data $Y = (y_i)$, we have log-likelihood

$$\ell(y, \lambda) = \sum_{i=1}^n \log \left(\frac{1}{Z_\lambda} \sum_x p(x, y_i) e^{\sum_j \lambda_j T_j(x, y_j)} \right)$$

and

$$\mathbb{E}_\lambda[T_k(x, y)] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\lambda[T_k(x, y_i)]$$

hence, we can find $\widehat{\lambda}$ by

1. *E-step:*
$$\widehat{T}_k^{(t)} = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\lambda(t)}[T_k(x, y_i \mid y_i)]$$
2. M-step: Find $\lambda(t+1)$ by
$$\mathbb{E}_{\lambda(t+1)}[T_k(x, y)] = \widehat{T}_k^{(t)}$$

MM Algorithm: The EM algoirhtm is a subset of a large class of *MM Algorithms* seeking to maximize $\ell(\theta)$ given $A(\theta, \tilde{\theta})$ satisfying

1. $A(\theta, \tilde{\theta}) \leq \ell(\theta)$
2. $A(\theta, \theta) = \ell(\theta)$

In this case, we define

$$\theta^{(t+1)} = \arg\max_\theta A(\theta, \theta^{(t)})$$

in which case $\ell(\theta^{(t)}) \leq \ell(\theta^{(t+1)})$