TODAY: Maximum Entropy

1. Note: No Homework Lecture Today! [new homework will be posted tomorrow]
2. Brandeis Dice
3. Maximum Entropy: general formulation
   - Examples
4. Exponential Families

Next Week: Maximum Entropy & MDL ; Connection to S-Values

Brandeis Dice (Jaynes 1957)

- $\mathcal{X} = \{1,2,\ldots,6\}$
- We found a strange looking die. We throw it 10000 times. We observe average nr of spots of 4.5 .
- Now we are asked to guess distribution of X. What should we do?
  - (1) we should perhaps set probs equal to freqs, but… we have not recorded all the frequencies!
  - (2) we pick the most uncertain one, which we take to be the one with Maximum Entropy, i.e.

$$P_{\text{me}} = \arg \max_{P : \mathbf{E}_{X \sim P}[X] = 4.5} H(P)$$

Brandeis Dice (Jaynes 1957)

- $\mathcal{X} = \{1,2,\ldots,6\}$
- We throw 10000 times. We observe average nr of spots of 4.5 .
- Now we are asked to guess distribution of X.
- We pick the most uncertain one, which we take to be the one with Maximum Entropy
  - Sounds like 'the least unreasonable one can do'
- How does the MaxEnt distribution look like? $(p_{\mathrm{me}}(1),\ldots,p_{\mathrm{me}}(6))$ = (0.05435,0.07877,0.11416,0.16545,0.23977,0.34749)

Brandeis Dice (Jaynes 1957)

- How does the MaxEnt distribution look like? $(p_{\mathrm{me}}(1),\ldots,p_{\mathrm{me}}(6))$ = (0.05435,0.07877,0.11416,0.16545,0.23977,0.34749)
  - Note that this doesn't have to be the true distribution!
  - P(X=4) = P(X=5) = 1/2 could be 'true', for example
- …so this distribution can never be more than a first rather wild guess
- …still, adopting the MaxEnt distribution may sometimes be reasonable

General Setting

- Suppose we want to make a prediction about a RV X
- If we know distribution of X, we can use that to make optimal predictions
- But here we deal with situation that we only have partial knowledge of distribution of X
  - knowledge of form: $P \in \mathcal{P}$ for **convex** $\mathcal{P}$
  - In lecture/book we only consider the special case of **linear constraints**, i.e. $\mathcal{P}$ of form $\mathcal{P} = \{P: E_P[\phi(X)] = t\}$ for some function $\phi: \mathcal{X} \to \mathbb{R}^k$

    (convex $\Rightarrow$ linear , but not vice versa)

General Setting

- We assume so many observations that we can safely set expectations to averages!
- dice problem: ϕ is identity! but in general, can be more complicated.
- According to Jaynes' maxent principle, we should pick the distribution in $\mathcal{P}$ maximizing entropy
  - dice example: distribution I just showed
  - More Realistic Examples: e.g. natural language processing, species modelling

The Good and The Bad

- Good Properties of MaxEnt procedure:
  - Unique solution: entropy is strictly concave
  - Uniformity: if consistent with constraint, will pick the uniform distribution [generalizes **Laplace's Principle of Insufficient Reason**]
  - If consistent with constraint, will pick distribution under which RVs are independent (ϕ = indicator functions)
  - For certain prediction problems, it gives minimax optimal predictions **(next week!)**
- Bad Properties:
  - Guess might be wrong **(Ex Nihilo Nihil!)**
  - …for other prediction problems, not at all 'optimal in any sense'

How to Compute MaxEnt Distributions

- Why do we get the answer we got?
- Let $\mathcal{P} = \{P: E_P[\phi(X)] = t\}$ for some function $\phi: \mathcal{X} \to \mathbb{R}^k$
- Let (T=transpose)

$$p_\beta(x) = \frac{1}{Z(\beta)} \cdot e^{\beta^T \phi(X)} \qquad Z(\beta) = \sum_{x \in \mathcal{X}} e^{\beta^T \phi(X)}$$

Theorem: suppose there exists $\tilde{\beta}$ s.t. $P_{\tilde{\beta}} \in \mathcal{P}$, i.e. $E_{X \sim P_{\tilde{\beta}}}[\phi(X)] = t$. Then:

$$P_{\tilde{\beta}} = P_{\text{me}} := \arg\max_{P \in \mathcal{P}} H(P)$$

Computing MaxEnt Distributions

$$p_\beta(x) = \frac{1}{Z(\beta)} \cdot e^{\beta^T \phi(X)} \qquad Z(\beta) = \sum_{x \in \mathcal{X}} e^{\beta^T \phi(X)}$$

Theorem: suppose there exists $\tilde{\beta}$ s.t. $p_{\tilde{\beta}} \in \mathcal{P}$ i.e.
$E_{X \sim P_{\tilde{\beta}}}[\phi(X)] = t$. Then: $P_{\tilde{\beta}} = P_{\text{me}} := \arg\max_{P \in \mathcal{P}} H(P)$

- Proof:

$$H(P) \leq \mathbf{E}_{X \sim P}[-\log P_{\tilde{\beta}}(X)] =$$
$$\mathbf{E}_{X \sim P}[-\tilde{\beta}^T \phi(X) + \log Z(\tilde{\beta})] = -\tilde{\beta}^T t + \log Z(\tilde{\beta}) =$$
$$\mathbf{E}_{X \sim P_{\tilde{\beta}}}[-\beta^T \phi(X) + \log Z(\tilde{\beta})] = H(P_{\tilde{\beta}})$$

- Strange (but correct) proof. We started by assuming the answer, and then showed that it must actually **be** the answer

Computing MaxEnt Distributions

$$p_\beta(x) = \frac{1}{Z(\beta)} \cdot e^{\beta^T \phi(X)} \qquad Z(\beta) = \sum_{x \in \mathcal{X}} e^{\beta^T \phi(X)}$$

Theorem: suppose there exists $\tilde{\beta}$ s.t. $p_{\tilde{\beta}} \in \mathcal{P}$ i.e. $E_{X \sim P_{\tilde{\beta}}}[\phi(X)] = t$. Then: $P_{\tilde{\beta}} = P_{\text{me}} := \arg\max_{P \in \mathcal{P}} H(P)$

- Usually constraints $\mathcal{P}$ are such that $\tilde{\beta}$ exists!
- Special case of "Boltzmann-Gibbs distribution" "maximum entropy distribution“ “exponential family”
  - arise frequently in physics
  - arise in statistics because they have finite-dimensional sufficient statistics (next week)

Example 1: Dice

$$p_\beta(x) = \frac{1}{Z(\beta)} \cdot e^{\beta \cdot X} \qquad Z(\beta) = e^{\beta} + e^{2\beta} + \ldots + e^{6\beta}$$

- Pick $\tilde{\beta}$ such that expectation is 4.5
- Note: as β ranges from $-\infty$ to ∞ , $E_{P_\beta}[X]$ ranges from 1 to 6

Example 2: Bernoulli

- $X = \{0,1\}$; $\mathcal{P} = \{P : E_P\,[X] = t\,\}$
- $P(X = 1) \cdot 1 + P(X = 0) \cdot 0 = t$
- i.e.
- $P(X = 1) = t$
- Note: as β ranges from $-\infty$ to ∞ , $E_{P_\beta}[X]$ ranges from 0 to 1 – the 'MaxEnt' model coincides with the Bernoulli model
- If you plug in $\beta = \log\left(\frac{p}{1-p}\right)$, you see that P_β is just Bernoulli distribution with mean p

Example 3: Independence if Consistent with Constraints

- $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2,\ \mathcal{X}_i = \{a, b\}$
- Constraint:
- $P(X_1 = a) = p$; $P(X_2 = a) = q$
- …rewrite as $E_P\left[1_{X_1=a}\right] = p$; $E_P\left[1_{X_2=a}\right] = q$
- $1_{X_1=a} = 1$ if $X_1 = a$; 0 otherwise.
- Solution must be of form
- $p_\beta(X_1, X_2) = \frac{1}{Z(\beta)} \cdot e^{\beta_1 1_{X_1=a} + \beta_2 1_{X_2=a}}$
- can be written as a product of something only dependent of X_1 and something only dependent of X_2 -> X_1 and X_2 must be independent under p_β

Example(s) 4, Continuous Data

- no constraints, $\mathcal{X} = [a, b] \Rightarrow$ MaxEnt is uniform distribution on $\mathcal{X}$
- $\mathcal{X} = \mathbb{R}^+$, constraint $E[X] = t \Rightarrow$ MaxEnt is exponential distribution with parameter $\frac{1}{t}$
- $\mathcal{X} = \mathbb{R}$, constraint $E[X] = \mu$, $\mathrm{var}[X] = \sigma^2 \Rightarrow$ MaxEnt is normal distribution with parameters μ,σ^2
  - [Reinterpretation of Central Limit Theorem: if we add and renormalize i.i.d. random variables [perform an operation that keeps μ, σ^2 the same] then the resulting distribution tends to the one with maximum entropy with this μ, σ^2]

The Good and The Bad, Revisited

- Good Properties of MaxEnt procedure:
  - Unique solution: entropy is strictly concave
  - Uniformity: if consistent with constraint, will pick the uniform distribution [generalizes **Laplace's Principle of Insufficient Reason**]
  - If consistent with constraint, will pick distribution under which RVs are independent (phi = indicator functions)
  - For certain prediction problems, it gives minimax optimal predictions **[next week!]**
- Bad Properties:
  - Guess might be wrong **(Ex Nihilo Nihil!)**
  - …for other prediction problems, not at all 'optimal in any sense'

General Setting

- Good Properties
- Bad Properties
- …they also simply arise in many practical situations, for different reasons [next week we'll see such a reason!]. So they are important to study even without the idea to use them as a 'first guess' of the underlying distribution

Exponential Families

- if $q(x) = 1$, then it is a 'maximum entropy' family

- Most models we have seen before are exponential families: Bernoulli, multinomial, normal, exponential, Gamma, Poisson, Pareto, Zipf, Beta, Gamma…: all exp families
- **…** also: Markov (need to extend definition to cover this),
- Gaussian (and other) Mixtures do not form an exponential family!

Sufficient Statistics!

- Why are exponential families easy to work with? Because they allow for **finite dimensional sufficient statistics (not depending on sample size)**
- **…and (with caveats) they are the only models with this property (Pitman-Koopman-Darmois)**
- "A sufficient statistic of a sample relative to a model summarizes *all* information in the sample that is important to make inferences relative to the model"

Sufficient Statistics!

- Sample size n: exponential families constructed by taking product distributions
- $p_\beta(x^n) = \frac{1}{Z(\beta)^n} e^{\beta^T \sum_{i=1..n} \phi(X_i)} \prod q(x_i)$

$$\max_\beta \log p_\beta(x^n) = \max_\beta \left(\beta \sum \phi(x_i) - n \log Z(\beta) + \sum \log q(x_i)\right)$$

- To determine this, you only need to know sum (equivalently, average) of ϕ **!**

Sufficient Statistics!

$$\max_{\beta} \log p_{\beta}(x^n) = \max_{\beta} \left(\beta \sum \phi(x_i) - n \log Z(\beta) + \sum \log q(x_i)\ \right)$$

- To determine this, you only need to know sum (equivalently, average) of ϕ **!**
- Bernoulli/binomial: need nr of 1s. <no other details>.
- Normal distribution: need mean and variance <no other details
- Poisson: only need mean

VERY easy to do statistics with: underlying reason why they are used so often. Not necessarily that they are good models of reality!

E.g **mixture models** do not have finite-dim suff stats.

Mean-Value Parameterization

- Theorem:

for every exponential family $\mathcal{M} = \{P_\beta : \beta \in \Theta_\beta\}$,
$E_{P_\beta}[\phi(X)]$ is strictly monotonically increasing as a function of β

  - [in the book this is also made precise for k-dim families with k > 1, i.e. $\phi: \mathcal{X} \to \mathbb{R}^k$, where it is not directly clear what 'monotonic' means]

- Intuition for proof: if β increases, then x with high $\phi(x)$ get exponentially more weight
- Therefore, we can identify a distribution in $\mathcal{M}$ by its mean of ϕ rather than the value of β

Mean-Value Parameterization

We can also identify a distribution in $\mathcal{M}$ by its mean of ϕ rather than the value of β. Thus we can always:

re-parameterize $\mathcal{M} = \{P_\beta : \beta \in \Theta_\beta\}$ as $\mathcal{M} = \{P_\mu : \mu \in \Theta_\mu\}$

where $\mu_\beta := E_{P_\beta}[\phi(X)]$

- ***β*: natural or canonical parameterization**
- ***μ*: mean-value parameterization**
- β_μ : inverse of μ_β
- Bernoulli: $\beta_\mu = \log \frac{\mu}{1-\mu}$; Exponential: $\beta_\mu = 1/\mu$
- Normal with mean 0, varying $\sigma^2 = E[X^2]$ mean (!)-value parameter: $\beta_{\sigma^2} = 1/(2\sigma^2)$

Nice Properties (“duality”)

- We have:
- $\mu_\beta = (\frac{d}{d\beta})\log Z(\beta)$ [multivariate: $\mu_\beta = \nabla \log Z(\beta)$]
- $\mathrm{var}_{P_\beta}(\phi) = \left(\frac{d^2}{d\beta^2}\right)\log Z(\beta) = I(\beta)$

  [multivariate: covariance matrix = Hessian= $I(\beta)$]

- …analogous properties for $\beta_\mu = (\frac{d}{d\mu})\log D(\mu || \mu_0)$
- $I(\mu) = \left(\frac{d^2}{d\mu^2}\right)\log D(\mu || \mu_0) = \frac{1}{I(\beta_\mu)} = \frac{1}{\mathrm{var}_{P_\mu}[\phi]}$

TODAY: Maximum Entropy

1. Note: No Homework Lecture Today! [new homework will be posted tomorrow]
2. Brandeis Dice
3. Maximum Entropy: general formulation
   - Examples
4. Exponential Families

Next Week: Maximum Entropy & MDL ; Connection to S-Values

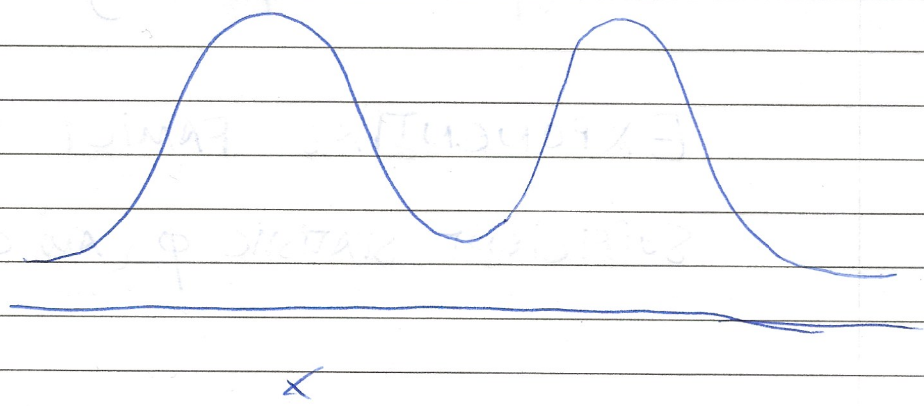
2 COMPONENT

GAUSSIAN MIXTURE

(NOT EXPONENTIAL FAMILY)

$$\left\{ P_{\alpha,\mu_1,\mu_2} : \alpha \in [0,1], \mu_1, \mu_2 \in \mathbb{R} \right\}$$

$$P_{\alpha,\mu_1,\mu_2}(x) = \alpha \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2}} + (1-\alpha)\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu_2)^2}{2}}$$



$\phi : \mathcal{X} \to \mathbb{R}^k \quad , \quad q : \mathcal{X} \to \mathbb{R}^+_0$

$\mathcal{M} = \{ P_\beta : \beta \in \Theta \} \quad , \Theta \subseteq \mathbb{R}^k$

$P_\beta(x) = \frac{1}{Z(\beta)} e^{\beta^T \phi(x)} q(x)$

$Z(\beta) = \int_{\mathcal{X}} e^{\beta^T \phi(x)} q(x) \, dx$

$\Theta = \{ \beta \in \mathbb{R}^k : Z(\beta) < \infty \}$

EXPONENTIAL FAMILY WITH

SUFFICIENT STATISTIC ϕ AND CARRIER q