Bayesian Methods in Machine Learning, Seminar: 1

Evgenii Egorov, egorov.evgenyy@ya.ru

Skoltech, 2020

MaxEnt

- We observe some data $p_e(x) = \frac{1}{N}\sum_{n=1}^{N}\delta(x - x_n)$.
  How should we select the probability density p to describe it?
- We could select some important quantities (feature mappings), that mean statistics describe our data:

$$\text{Mapping: } \phi_\alpha : \mathcal{X} \to \mathbb{R}, \alpha \in I, \text{where } \alpha \text{ could be both: discrete or continuous,}$$

$$\text{Important statistics: } \mu_\alpha = \langle \phi_\alpha \rangle_p, \text{ we observe : } \hat{\mu}_\alpha = \langle \phi_\alpha \rangle_{p_e} = \frac{1}{N}\sum_{n=1}^{N}\phi_\alpha(x_n).$$

- We don't have preferences and would like to have smooth model, so we would like to maximise the entropy:

$$\max_{p \in \mathcal{P}} H[p], \text{ st } \hat{\mu}_\alpha = \langle \phi_\alpha \rangle_p, \quad \forall \alpha \in I.$$

Our problem:

$$\max_{p \in \mathcal{P}} H[p], \text{ st } \hat{\mu}_\alpha = \langle \phi_\alpha \rangle_p, \quad \forall \alpha \in I.$$

Consider the Lagrangian and find optimal p, which depends on dual variables.
Hint:

$$\frac{\partial}{\partial p} \int_{\mathcal{X}} p \log p \; dx = \log p + 1.$$

MaxEnt: Solution

Solution:

$$\begin{aligned}
&\mathcal{L} = -\langle \log p \rangle_p + \lambda^T \left(\langle \phi \rangle_p - \hat{\mu} \right) + \mu \left(\langle 1 \rangle_p - 1 \right), \\
&\frac{\partial}{\partial p} \mathcal{L} = -(\log p + 1) + \lambda^T \phi + \mu, \\
&\boxed{p \propto \exp\{\langle \lambda, \phi(x) \rangle\}}.
\end{aligned}$$

A bit heuristic derivation of deviation:

$$\int (p + \varepsilon h) \log(p + \varepsilon h) \; dx - \int p \log p \; dx = \varepsilon \int h(\log p + 1) \; dx + o(\varepsilon).$$

MaxEnt: Problem

Let's prove the uniqueness of the solution.
Plan:

- Consider the another solution, i.e. maximum entropy with satisfying moment constrains
- Proof that

Hint:

$$- D_{KL}[p|q] = \int p \log \tfrac{q}{p} \, dx \leq \{\text{by Jensen inequality}\} \leq \log \int p \frac{q}{p} \, dx = 0.$$

Solution:
Consider we are given two solutions: p and q. Let's compare their entropy.

$$H[q] = -\int q \log q \; dx = -\int q \log \frac{q}{p} \; dx - \int q \log p \; dx =$$

$$- D_{KL}[q|p] - \int q \left[\langle \lambda, \phi(x) \rangle - \log Z(\lambda)\right] \; dx =$$

$$- D_{KL}[q|p] - \int p \left[\langle \lambda, \phi(x) \rangle - \log Z(\lambda)\right] \; dx =$$

$$- D_{KL}[q|p] + H[p] \leq H[p].$$

$$\boxed{H[q] \leq H[p]}.$$

MaxEnt

Given value of the **mean parameters** $\hat{\mu}_\alpha = \langle \phi_\alpha \rangle$, we obtain distribution:

$$p(x;\lambda) = \exp\left\{\langle \lambda, \phi(x) \rangle - A(\lambda)\right\}, \;\; A(\lambda) = \log \int \exp\left\{\langle \lambda, \phi(x) \rangle\right\} \; dx.$$

We could assume further:

- $\phi(x)$ is the minimal sufficient statistic if $\nexists: \;\; \lambda \neq 0, \; \langle \phi(x), \lambda \rangle = \text{const}.$
- Space of the **natural parameters**: $\Omega = \{\lambda \in \mathbb{R}^d | A(\lambda) < +\infty\}$.
- Space of the **mean parameters**: $\mathcal{M} = \{\mu \in \mathbb{R}^d | q : \mu = \langle \phi \rangle_q\}$.

What is the correspondence between the **natural parameters** and **mean parameters**?

Mapping between parameters $A(\lambda)$: Problem

$$p(x;\lambda) = \exp\left\{\langle \lambda, \phi(x)\rangle - A(\lambda)\right\}, \;\; A(\lambda) = \log \int \exp\left\{\langle \lambda, \phi(x)\rangle\right\} \; dx.$$

We would like to find mapping between natural and mean parameters. In order to do this, consider two problems:

- $\nabla A(\lambda) = \dots$
- $\lambda = \arg\max\limits_{\lambda} \sum\limits_{n=1}^{N} \frac{1}{N} \log p(x_n;\lambda) = \dots$

Solution

From natural parameters to mean:

$$\nabla A(\lambda) = \int \phi(x) \exp\{\langle \lambda, \phi(x) \rangle - A(\lambda)\rangle\} \, dx = \boxed{\langle \phi \rangle}.$$

From mean parameters to natural:

$$\max_{\lambda} \sum_{n=1}^{N} \frac{1}{N} \log p(x_n; \lambda) = \max_{\lambda} \left\langle \lambda, \frac{1}{N} \sum_{n=1}^{N} \phi(x_n) \right\rangle - A(\lambda).$$

$$\boxed{\frac{1}{N} \sum_{n=1}^{N} \phi(x_n) = \nabla A(\lambda)}.$$

The reverse mapping is the optimization problem. It is one-to-one, if the optimization problem is concave.

Mapping between parameters $A(\lambda)$: Problem

$$p(x;\lambda) = \exp\left\{\langle \lambda, \phi(x) \rangle - A(\lambda)\right\}, \quad A(\lambda) = \log \int \exp\left\{\langle \lambda, \phi(x) \rangle\right\} \, dx.$$

Let's establish the convexity of the $A(\lambda)$:

$$\nabla_{\lambda\lambda} A(\lambda) = \dots$$

Mapping between parameters $A(\lambda)$: Solution

$$p(x;\lambda) = \exp\left\{\langle \lambda, \phi(x)\rangle - A(\lambda)\right\}.$$

$$\nabla A(\lambda) = \int \phi(x)\exp\{\langle \lambda, \phi(x)\rangle - A(\lambda)\rangle\}\ dx = \boxed{\langle \phi \rangle}.$$

Solution:

$$\nabla_{\lambda\lambda} A(\lambda) = \int \phi(x) \nabla_\lambda p(x;\lambda)\ dx = \int \phi(x) p(x;\lambda) \nabla_\lambda \log p(x;\lambda)\ dx.$$

$$\nabla_\lambda \log p(x;\lambda) = \nabla_\lambda\left[\langle \lambda, \phi(x)\rangle - A(\lambda)\right] = \phi(x) - \nabla A(\lambda) = \phi(x) - \langle \phi \rangle.$$

$$\int p(x;\lambda)\phi(x)[\phi(x) - \langle \phi \rangle]^T\ dx = \langle \phi \phi^T \rangle_p - \langle \phi \rangle_p^2 = \boxed{\mathrm{Cov}(\phi) \succ 0}.$$

Mapping between parameters $A(\lambda)$

As $A(\lambda)$ is the convex function, we could consider its fenchel conjugate:

$$A^*(\mu) = \sup_{\lambda \in \Omega} \langle \lambda, \mu \rangle - A(\lambda)$$

$$\text{Recall MLE problem:} \lambda = \arg\max_{\lambda \in \Omega} \langle \lambda, \frac{1}{N} \sum_{n=1}^{N} \phi(x_n) \rangle - A(\lambda)$$

$$A^*(\mu) = \langle \lambda(\mu), \mu \rangle - A(\lambda(\mu)) = -H[p(x; \lambda(\mu))].$$

Next time we continue investigate properties of the exponential family and natural-mean parametrization.

Problem: Heads/Tail Probability Inference

Consider following model:

$$p(\theta|\tau) = \frac{\Gamma(\tau_1 + \tau_2))}{\Gamma(\tau_1)\Gamma(\tau_2)}\theta^{\tau_1 - 1}(1-\theta)^{\tau_2}, \tau > 0 \quad p(x|\theta) = \theta^x(1-\theta)^{1-x},\ x \in \{0, 1\}, \theta \in (0, 1)$$

Observed $X = (x_1, \ldots x_N)$, find:

- MLE
- $p(\theta|X, \tau)$, expectation, MAP
- Predictive distribution

Solution: MLE

$$\theta^{MLE} = \arg\max_\theta \prod_{n=1}^{N} p(x_n|\theta) = \arg\max_\theta \sum_{n=1}^{N} \log p(x_n|\theta)$$

$$\log p(X|\theta) = \left[\sum_{n=1}^{N} x_n\right] \log\theta + \left[N - \sum_{n=1}^{N} x_n\right] \log(1-\theta)$$

$$\nabla_\theta \log p(X|\theta) = \left[\frac{1}{\theta}\overline{x} - \frac{1}{1-\theta}(1-\overline{x})\right] N = 0$$

$$\theta^{MLE} = \overline{x}$$

Recall, that for exponential family

$$\log p(x_n|\lambda) = \langle\theta, T(X)\rangle - F(\theta)$$

$$F(\theta) = \int_\Theta \exp\left(\langle\theta, T(X)\rangle\right) d\mu(x)$$

Problem is the convex optimization problem.

Solution: Posterior density

Bayes rule:

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{\int\limits_{\Theta} p(X|\theta)p(\theta)d\theta}$$

$$p(\theta|X,\tau) = \frac{1}{Z}p(X|\theta)p(\theta|\tau) \propto \left(\prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{1-x_n}\right)\theta^{\tau_1-1}(1-\theta)^{\tau_2-1}$$

$$p(\theta|X,\tau) \sim Beta\left(\tau_1 + \sum_{n=1}^{N} x_n, \tau_2 + N - \sum_{n=1}^{N} x_n\right)$$

Note, that we have as the posterior same distribution as the prior, with easy incremental update of the parameters.

Solution: Point estimators from $p(\theta|X, \tau)$

$$\langle\theta\rangle_{p(\theta|X,\tau)} = \frac{\sum_{n=1}^{N} x_n + \tau_1}{N + \tau_1 + \tau_2} = \left(\frac{\tau_1 + \tau_2}{N + \tau_1 + \tau_2}\right) \frac{\tau_1}{\tau_1 + \tau_2} + \left(1 - \frac{\tau_1 + \tau_2}{N + \tau_1 + \tau_2}\right) \overline{x}$$

$$\langle\theta\rangle_{p(\theta|X,\tau)} = \alpha\langle\theta\rangle_{p(\theta)} + (1 - \alpha)\theta^{MLE}$$

Convex combination of prior and MLE estimators. Moreover, as $N \to \infty$

$$\langle\theta\rangle_{p(\theta|X,\tau)} \to \theta^{MLE}, \ \mathbb{D}_{p(\theta|X,\tau)}\theta \to 0$$

$$\theta^{\text{MAP}} = \frac{\sum_{n=1}^{N} x_n + \tau_1 - 1}{N + \tau_1 + \tau_2 - 2}$$

Solution: Predictive Distribution

$$p(x^*|X) = \int_{\Theta} p(x^*|\theta)p(\theta|X,\tau)d\theta$$

(We have here the assumption: $x_{new} \perp X|\theta$ here.)

$$p(x^* = 1|X) = \int\limits_{\Theta} \theta^{x^*}(1-\theta)^{1-x^*}\frac{\Gamma(\tau_1^{'}+\tau_2^{'})}{\Gamma(\tau_1^{'})\Gamma(\tau_2^{'})}\theta^{\tau_1^{'}-1}(1-\theta)^{\tau_2^{'}-1}d\theta =$$

$$= \frac{\Gamma(\tau_1^{'}+\tau_2^{'})}{\Gamma(\tau_1^{'})\Gamma(\tau_2^{'})}\int\limits_{\Theta} \theta^{x^*+\tau_1^{'}-1}(1-\theta)^{\tau_2^{'}-x^*}d\theta = \frac{Z_{\text{update}}}{Z_{\text{posterior}}} = \frac{\sum\limits_{n=1}^{N} x_n + \tau_1}{N + \tau_1 + \tau_2}$$

Solution: Simulation

Consider some **frequentist** simulation study:

- $X \sim \text{Binomial}(p_{true})$
- Update θ^{MLE}, $p(\theta|X, \tau)$

Simulation Study: **Seminar 1 - BetaAnimation.ipynb**

Conjugate Prior Construction

We obtain nice results with conjugate prior and likelihood:

- posterior distribution is the same distribution as prior, with additive updates of the parameters
- predictive distribution has analytic form

So, how should we construct prior distribution, to make it conjugate to our model?

Conjugate Prior Construction: Natural

Consider our model from exponential family:

$$p(x|\eta) = \exp\left(\langle \eta, T(x) \rangle - F(\eta)\right)$$

Then, as likelihood under iid $X = (x_1, \dots, x_n)$:

$$p(X|\eta) = \exp\left(\langle \eta, \sum_{n=1}^{N} T(x_i) \rangle - NF(\eta)\right)$$

Now we just write prior density at the same form:

$$p(\eta|\tau, n_0) = H(\tau, n_0) \exp\left(\langle \eta, \tau \rangle - n_0 F(\eta)\right), \;\; n_0 > 0$$

Note, that here $H(\tau, n_0)$ is normalizing factor! and $F(\eta)$ is statistics!

Conjugate Prior Construction: Natural

$$p(\eta|X,\tau,n_0) \propto p(X|\eta)p(\eta|\tau,n_0) \propto \exp\left(\langle\eta, \tau + \sum_{n=1}^{N} T(x_i) - (n_0+N)F(\eta)\right)$$

Hence, posterior is nothing more than p$(\eta|\tau', n_0')$:

$$\tau' = \tau + \sum_{i=1}^{N}$$

$$n_0' = n_0 + N$$

Problem: Exponential Family Predictive Distribution

Consider model:

$$p(x|\eta) = \exp\left(\langle \eta, T(x)\rangle - F(\eta)\right)$$

And prior:

$$p(\eta|\tau, n_0) = H(\tau, n_0)\exp\left(\langle \eta, \tau\rangle - n_0 F(\eta)\right), \;\; n_0 > 0$$

After observation $X =_{\text{iid}} (x_1, \dots, x_N)$

Find:

$$p(x_{\text{new}}|X) = \dots?$$

Solution: Exponential Family Predictive Distribution

$$p(x_*|X) = \int p(x^*|\eta)p(\eta|X,\tau,n_0)d\eta =$$

$$= \int \exp\left(\langle \eta, T(x^*) \rangle - F(\eta)\right) H(\tau', n_0') \exp\left(\langle \eta, \tau' \rangle - n_0' F(\eta)\right) d\eta =$$

$$= H(\tau', n_0') \int \exp\left(\langle \eta, T(x^*) + \tau' \rangle - (1 + n_0')F(\eta)\right) = \frac{H(\tau', n_0')}{H(\tau' + T(x^*, n_0' + 1)} =$$

$$= \frac{H(\tau + \sum_{n=1}^{N} T(x_n), n_0 + N)}{H(\tau + \sum_{n=1}^{N} T(x_n) + T(x^*), n_0 + N + 1)}$$