Bayesian Methods in Machine Learning, Seminar: 3

Evgenii Egorov, egorov.evgenyy@ya.ru

Skoltech, 2020

Recap: 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:

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

Recap: MaxEnt $\to$ Exponential Family

MaxEnt solution: $p(x;\lambda) \propto \exp(\langle \phi(x), \lambda \rangle)$.

An **exponential family** is a set of probability distributions admitting the following **canonical decomposition**:

- $p(x;\lambda) = \exp\left(\langle \phi(x), \lambda \rangle - A(\lambda) + k(x)\right)$:
- $\phi(x)$ is the minimal sufficient statistic if $\nexists : \ \lambda \neq 0, \ \langle \phi(x), \lambda \rangle = \text{const}$.
- $\langle\rangle$ is the corresponding inner product
- $A(\lambda) = \log \int \exp\left(\langle \phi(x), \lambda \rangle + k(x)\right) \ dx$ is the log-normalizer

$k(x)$ is the carrier measure, usually corresponds to the Lebesgue or Counting.

Long list, what is **important for us**:

- Decomposition on the parameter-dependent and "data"-dependent functions
- Linear (**inner product**) interaction between this parts.

Recap: RVM Regression Model

For data:

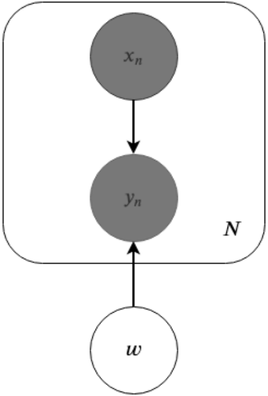
$$x_n \in \mathbb{R}^D, w \in \mathbb{R}^D, t_n \in \mathbb{R},$$
$$(X, \mathbf{t}) = \{(x_n, t_n)\}_{n=1}^N.$$

Consider following model:

$$p(t_n|x_n, w; \beta) = \mathcal{N}(t_n|\mathbf{w}^T x_n, \beta^{-1}),$$

$$p(\mathbf{t}|X, \mathbf{w}; \beta) = \prod_{n=1}^{N} p(t_n|x_n, \mathbf{w}; \beta) = \mathcal{N}(\mathbf{t}|X\mathbf{w}, \beta^{-1} I_{N\times N}),$$

$$p(\mathbf{w}; \alpha) = \prod_{d=1}^{D} \mathcal{N}(w_d|0, \alpha_d^{-1}) = \mathcal{N}(\mathbf{w}|0, A^{-1}).$$



Recap: RVM Regression Model

We could note, that the posterior $p(\mathbf{w}|(X, \mathbf{t}))$ is closed-form, i.e. Normal distribution:

$$\log p(\mathbf{w}|(X, \mathbf{t})) \propto^{+} \underbrace{-\frac{\beta}{2}(\mathbf{t} - X\mathbf{w})^T(\mathbf{t} - X\mathbf{w}) - \frac{1}{2}\mathbf{w}^T A \mathbf{w}}_{\text{Quadratic function over } \mathbf{w}}.$$

We can also get the marginal distribution in the form also:

$$p(\mathbf{t}|X) = |2\pi\beta^{-1}|^{-\frac{N}{2}}|2\pi A^{-1}|^{-\frac{1}{2}} \exp(f(\mathbf{w}^*))|2\pi[\beta X^T X + A]^{-1}|,$$
$$f(\mathbf{w}) = f(\mathbf{w}^*) - \tfrac{1}{2}(\mathbf{w} - \mathbf{w}^*)^T[\beta X^T X + A](\mathbf{w} - \mathbf{w}^*).$$

Normal Likelihood + Normal prior = Closed form equations.

General Recipe? **Conjugate prior.**

Problem: Familiar Distributions as Members of Exponential Familiy

Canonical representation:

- $p(x;\lambda) = \exp\left(\langle \phi(x), \lambda\rangle - A(\lambda) + k(x)\right)$:
- $\phi(x)$ is the minimal sufficient statistic if $\nexists: \ \lambda \neq 0, \ \langle \phi(x), \lambda \rangle = \text{const}$.
- $\langle\rangle$ is the corresponding inner product
- $A(\lambda) = \log \int \exp\left(\langle \phi(x), \lambda\rangle + k(x)\right) \ dx$ is the log-normalizer

Problem: Derive canonical representation for the following members of the exponential family:

- Normal: $(2\pi\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$,
- Multinomial: $\dfrac{N!}{x_1!x_2!\ldots x_K!}\pi_1^{x_1}\pi_2^{x_2}\ldots\pi_K^{x_K}$, $\sum_{k=1}^K x_k = N$, $\sum_{k=1}^K \pi_k = 1$.

Solutions: Normal Distribution as Member of Exponential Family

Normal distribution:

$$\exp\left(-\frac{1}{2}\mathrm{Tr}[xx^T - 2\mu x^T + \mu\mu^T]\Sigma^{-1} - \frac{1}{2}\log\|\Sigma\| - \frac{d}{2}\log 2\pi\right) =$$
$$= \exp\left(\mathrm{Tr}(-\frac{1}{2}xx^T\Sigma^{-1}) + \mathrm{Tr}(x^T\Sigma^{-1}\mu) - \frac{1}{2}\mathrm{Tr}(\mu\mu^T\Sigma^{-1}) + \frac{1}{2}\log\|\Sigma\|^{-1} - \frac{d}{2}\log 2\pi\right).$$

- $\phi(x) = (x, -xx^T)$,
- $\lambda = (\Sigma^{-1}\mu, \frac{1}{2}\Sigma^{-1})$.
- $A(\lambda) = \frac{1}{2}\mathrm{Tr}(\mu\mu^T\Sigma^{-1}) - \frac{1}{2}\log\|\Sigma\|^{-1} + \frac{d}{2}\log 2\pi, \quad \mu = \frac{1}{2}\Lambda_2^{-1}\lambda_1,\ \Sigma = \frac{1}{2}\lambda_2^{-1}$.
  Hence: $A(\lambda) = \frac{1}{4}\mathrm{Tr}(\Lambda_2^{-1}\lambda_1\lambda_1^T) - \frac{1}{2}\log\||\lambda_2\| + \frac{d}{2}\log\pi$.
- $k(x) = 0$.

Solutions: Multinomial Distribution as Member of Exponential Family

Multinomial Distribution:

$$\exp\left(\sum_{k=1}^{K} x_k \log \pi_k\right) = \{\text{Minimality!}\} =$$

$$= \exp\left(\sum_{k=1}^{K-1} x_k \log \frac{\pi_k}{1-\sum_{k=1}^{K-1}\pi_k} + N\log(1-\sum_{k=1}^{K-1}\pi_k)\right).$$

- $\phi(x) = x_k,\ \ k = 1, \ldots, K-1.$
- $\lambda_k = \log \frac{\pi_k}{1-\sum_{k=1}^{K-1}\pi_k},\ \ k = 1, \ldots, K-1$
- $A(\lambda) = N\log(1-\sum_{k=1}^{K}\pi_k) - \log N! =$

  $$\pi_k = \frac{\exp(\lambda_k)}{\sum_{k=1}^{K-1}\exp(\lambda_k)} + \frac{1}{1+\sum_{k=1}^{K-1}\exp(\lambda_k)} = \text{Soft-Max}(\lambda_k), \lambda_K = 0$$

  .
- $k(x) = -\sum_{k=1}^{K}\log x_k!.$

Problem: Heads/Tail Probability Inference

Consider following model:

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

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

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

$$\log p(x_n|\lambda) = \langle\theta, \phi(x_n)\rangle - A(\theta).$$

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

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 $N \to \infty$

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

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

Solution: Predictive Distribution

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.)

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

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

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:

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

Then, as likelihood under iid $X = (x_1, \ldots, x_N)$:

$$p(X|\lambda) = \exp\left(\langle \lambda, \sum_{n=1}^{N} \phi(x_n) \rangle - NA(\lambda)\right).$$

Now we just write prior density at the same functional form:

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

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

Conjugate Prior Construction: Natural

Likelihood $\times$ Prior:

$$p(\lambda|X, \tau, n_0) \propto \exp\left(\langle\lambda, \sum_{n=1}^{N}\phi(x_n)\rangle - NA(\lambda)\right)\exp\left(\langle\lambda, \tau\rangle - n_0 A(\lambda)\right),$$

$$p(\lambda|X, \tau, n_0) \propto p(X|\lambda)p(\lambda|\tau, n_0) \propto \exp\left(\langle\lambda, \tau + \sum_{n=1}^{N}\phi(x_n)\rangle - (n_0 + N)A(\lambda)\right).$$

Hence, posterior is nothing more than $\mathrm{p}(\lambda|\tau', n_0')$:

$$\tau' = \tau + \sum_{n=1}^{N}\phi(x_n),$$

$$n_0' = n_0 + N.$$

Problem: Exponential Family Predictive Distribution

Consider model:

$$p(x|\lambda) = \exp\left(\langle\lambda, \phi(x)\rangle - A(\lambda)\right),$$

And prior:

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

.

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

$$p(x^*|X) = \ldots?$$

Solution: Exponential Family Predictive Distribution

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

$$= \int \exp\left(\langle\lambda,\phi(x^*)\rangle - A(\lambda)\right) H(\tau',n_0') \exp\left(\langle\lambda,\tau'\rangle - n_0' A(\lambda)\right) d\lambda =$$

$$= H(\tau',n_0') \int \exp\left(\langle\lambda,\phi(x^*)+\tau'\rangle - (1+n_0')A(\lambda)\right) = \frac{H(\tau',n_0')}{H(\tau'+\phi(x^*,n_0'+1)} =$$

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

Problem: the Posterior mean as Convex Combination

Consider model:

$$p(x|\lambda) = \exp\left(\langle \lambda, \phi(x) \rangle - A(\lambda)\right),$$

And prior:

$$p(\lambda|\tau, n_0) = H(\tau, n_0) \exp\left(\langle \lambda, \tau \rangle - n_0 A(\lambda)\right), \quad n_0 > 0$$

.
After observation $X =_{\text{iid}} (x_1, \ldots, x_N)$,
Find:

$$\langle \mu(\lambda) | X, \tau, n_0 \rangle = ?$$