Bayesian Methods in Machine Learning
Seminar: Bayesian Linear Regression ARD and Sequential Updates

Evgenii Egorov, Evgenii.Egorov@skoltech.ru

Skoltech

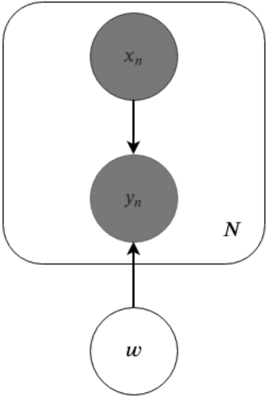
Relevance Vector Machine: Regression Model

For data:

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

Consider following model:

$$p(t_n|x_n, w; \beta) = \mathcal{N}(t_n|\mathbf{w}^T x_n, \beta^{-1}),$$
$$p(\mathbf{t}|X, \mathbf{w}; \beta) = \prod_{n=1}^N p(t_n|x_n, \mathbf{w}; \beta) = \mathcal{N}(\mathbf{t}|X\mathbf{w}, \beta^{-1}),$$
$$p(\mathbf{w}; \alpha) = \prod_{d=1}^D \mathcal{N}(w_d|0, \alpha_d^{-1}) = \mathcal{N}(\mathbf{w}|0, A^{-1}).$$



RVM: Posterior derivation

Consider for now, that we are given parameters of the prior α and noise β. Let's derive posterior over the w.

$$p(\mathbf{w}|(X,\mathbf{t})) = \frac{1}{Z}\underbrace{\mathcal{N}(\mathbf{t}|X\mathbf{w},\beta^{-1})}_{\text{Likelihood}}\underbrace{\mathcal{N}(\mathbf{w}|0,A^{-1})}_{\text{Prior}}.$$

We could note, that $p(\mathbf{w}|(X,\mathbf{t}))$ is Normal distribution:

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

- Hence, we could find **expectation** and **covariance** to define p(**w**|(X,**t**))
- For Normal distribution: **expectation** is the mode and inverse Hessian of log-density is **covariance**.

Normal Distribution: mode, expectation, covariance, hessian

Consider the normal distribution:

$$\mathcal{N}(y|\mu,\Sigma) = |2\pi\Sigma|^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu))$$

Let's find its mode. Hence, consider the $\frac{\partial}{\partial y}\log\mathcal{N}(y|\mu,\Sigma) = 0$:

$$\frac{\partial}{\partial y}\log\mathcal{N}(y|\mu,\Sigma) = -\frac{1}{2}\frac{\partial}{\partial y}(y-\mu)^T\Sigma^{-1}(y-\mu) = -\Sigma^{-1}(y-\mu) \Rightarrow \boxed{y^* = \mu}.$$

Let's find $\frac{\partial}{\partial^2 y}\log\mathcal{N}(y|\mu,\Sigma)$:

$$\frac{\partial^2}{\partial_{ij} y}\log\mathcal{N}(y|\mu,\Sigma) = \frac{\partial}{\partial y} - \Sigma^{-1}(y-\mu) = -\Sigma^{-1} \Rightarrow \boxed{-[\frac{\partial^2}{\partial_{ij} y}\log\mathcal{N}(y|\mu,\Sigma)]^{-1} = \Sigma}.$$

Hence, we could easily find moments of Normal distribution given its unnormilized density.

RVM: Posterior derivation

We could note, that $p(\mathbf{w}|(X, \mathbf{t}))$ is Normal distribution:

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

Let's find the expectation:

$$\begin{aligned} &\frac{\partial}{\partial \mathbf{w}} \log p(\mathbf{w}|(X, \mathbf{t})) = 0 \\ &-\beta X^T(X\mathbf{w} - \mathbf{t}) - A\mathbf{w} = 0 \\ &\mathbf{w}^* = \boxed{\beta[\beta X^T X + A]^{-1} X^T \mathbf{t} = \mu_w}. \end{aligned}$$

Let's find the covariance:

$$\frac{\partial^2}{\partial_{ij}\mathbf{w}} \log p(\mathbf{w}|(X, \mathbf{t})) = \frac{\partial}{\partial \mathbf{w}}[-\beta X^T(X\mathbf{w} - \mathbf{t}) - A\mathbf{w}] = -\beta X^T X - A \Rightarrow \boxed{[\beta X^T X + A]^{-1} = \Sigma_w}.$$

RVM: Sequential Updates of the Posterior

Consider the simplest sequential case: our data is **i.i.d**. Then we could note the following:

$$p(\mathbf{w}|(X,\mathbf{t})^1,(X,\mathbf{t})^2) \propto \underbrace{p(\mathbf{w})p((X,\mathbf{t})^1|\mathbf{w})}_{\text{Posterior after first data chunk}} p((X,\mathbf{t})^2|\mathbf{w}) \propto \underbrace{p(\mathbf{w}|(X,\mathbf{t})^1)}_{\text{New prior is "old" posterior}} p((X,\mathbf{w}^2)|\mathbf{w}).$$

Hence, we could store only the parameters of the model and drop data form previous steps. Let's derive equations for our mode. It will differ only slightly from previous derivations, as now our prior has non-zero mean.

$$\log p(\mathbf{w}|(X,\mathbf{t})^1,(X,\mathbf{t})^2) \propto -\frac{\beta}{2}(\mathbf{t}^{(2)} - X^{(2)}\mathbf{w})^T(\mathbf{t}^{(2)} - X^{(2)}\mathbf{w}) - \tfrac{1}{2}(\mathbf{w} - \mu_w)^T\Sigma_w^{-1}(\mathbf{w} - \mu_w).$$

And we need to do the same things: find mean and covariance using knowledge about mode and hessian.

RVM: Sequential Updates of the Posterior

$$\log p(\mathbf{w}|(X,\mathbf{t})^1,(X,\mathbf{t})^2) \propto -\frac{\beta}{2}(\mathbf{t}^{(2)} - X^{(2)}\mathbf{w})^T(\mathbf{t}^{(2)} - X^{(2)}\mathbf{w}) - \tfrac{1}{2}(\mathbf{w}-\mu_w)^T\Sigma_w^{-1}(\mathbf{w}-\mu_w).$$

Let's find the expectation:

$$\begin{aligned}
&\tfrac{\partial}{\partial \mathbf{w}} \log p(\mathbf{w}|(X,\mathbf{t})^{1,2}) = 0 \\
&-\beta X_{(2)}^T(X_{(2)}\mathbf{w} - \mathbf{t}_{(2)}) - \Sigma_w^{-1}(\mathbf{w}-\mu_w) = 0 \\
\mathbf{w}^* &= \boxed{[\beta X_{(2)}^T X_{(2)} + \Sigma_w^{-1}]^{-1}(\beta X_{(2)}^T \mathbf{t}_{(2)} + \Sigma_w^{-1}\mu_w) = \mu_w^{(1,2)}}
\end{aligned}$$

Let's find the covariance:

$$\tfrac{\partial^2}{\partial_{ij}\mathbf{w}} \log p(\mathbf{w}|(X,\mathbf{t})) = \tfrac{\partial}{\partial \mathbf{w}}[-\beta X_{(2)}^T(X_{(2)}\mathbf{w} - \mathbf{t}_{(2)}) - \Sigma_w^{-1}\mathbf{w}] \Rightarrow \boxed{[\beta X_{(2)}^T X_{(2)} + \Sigma_w^{-1}]^{-1} = \Sigma_w^{(1,2)}}.$$

Sequential Updates: Practise in Python

So, we start with the prior: $\mathcal{N}(\mathbf{w}|0, A)$ and get:

- Observe $(X, \mathbf{t})^1$ and obtain:

$$\Sigma_w^{-1} = [\beta X_{(1)}^T X_{(1)} + A]$$
$$\mu_w = \Sigma_w \beta X_{(1)}^T \mathbf{t}_{(1)}$$

- Observe $(X, \mathbf{t})^2$ and then for joint data:

$$\Sigma_{w(1,2)}^{-1} = [\beta X_{(2)}^T X_{(2)} + \Sigma_w^{-1}]$$
$$\mu_{w(1,2)} = \Sigma_{w(1,2)}(\beta X_{(2)}^T \mathbf{t}_{(2)} + \Sigma_w^{-1}\mu_w)$$

- . . .
  Let's open the notebook in colab and implement this [click]

Relevance Vector Machine: Evidence Optimization

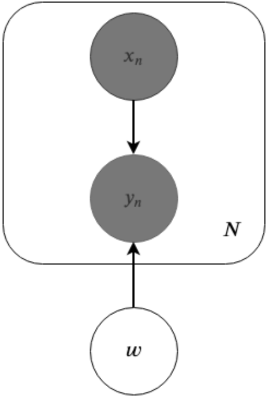
Let's recall our model.

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

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

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

How should we select the values α, β?



Relevance Vector Machine: Evidence Optimization

Max Evidence of the model:

$$\max_{\alpha,\beta} \log p(\mathbf{t}|X) = \max_{\alpha,\beta} \log \int \mathcal{N}(\mathbf{t}|X\mathbf{w}, \beta^{-1}I)\mathcal{N}(\mathbf{w}|0, A^{-1})d\mathbf{w}.$$

Let's take the integral:

$$p(\mathbf{t}|X) = |2\pi\beta^{-1}|^{-\frac{N}{2}}|2\pi A^{-1}|^{\frac{1}{2}} \int \exp\left(-\frac{\beta}{2}\|X\mathbf{w} - \mathbf{t}\|_2^2 - \frac{1}{2}\mathbf{w}^T A\mathbf{w}\right) d\mathbf{w}.$$

Again, we would use the simple plan:

- We know that integral over quadratic energy is the normalizing constant of Normal distribution
- Play a bit with mode, hessian and etc.

Let's do this.

Relevance Vector Machine: Evidence Optimization

Consider the function:

$$f(\mathbf{w}) = -\frac{\beta}{2}\|X\mathbf{w} - \mathbf{t}\|_2^2 - \frac{1}{2}\mathbf{w}^T A\mathbf{w}.$$

We already derived its extremum $\mathbf{w}^*$ (5) and hessian (5). So, using second-order Taylor expression, which is exact for quadratic functions, we obtain:

$$f(\mathbf{w}) = f(\mathbf{w}^*) - \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^T[\beta X^T X + A](\mathbf{w} - \mathbf{w}^*).$$

Now we can get the value of the integral:

$$p(\mathbf{t}|X) = |2\pi\beta^{-1}|^{-\frac{N}{2}}|2\pi A^{-1}|^{\frac{1}{2}}\exp(f(\mathbf{w}^*))\int \underbrace{\exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^T[\beta X^T X + A](\mathbf{w} - \mathbf{w}^*)\right) d\mathbf{w}}_{\text{Normalizing constant of the Normal distribution}} =$$

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

Relevance Vector Machine: Evidence Optimization

Finally, our optimisation problem is:

$$\arg\max_{\beta,\alpha} \log p(\mathbf{t}|X) = \arg\max_{\beta,\alpha} \frac{N}{2}\log\beta + \frac{1}{2}\log|A| - \frac{1}{2}\log|\beta X^T X + A| + f(\mathbf{w}^*)$$

However the extremum point $\mathbf{w}^*$ its self depends on the α, β. To simplify the problem, we introduce the lower bound insted:

$$\boxed{\frac{N}{2}\log\beta + \frac{1}{2}\log|A| - \frac{1}{2}\log|\left(\beta X^T X + A\right)| - \frac{\beta}{2}\|\mathbf{t} - X\mu\|_2^2 - \frac{1}{2}\mu^T A\mu}.$$

Now, we should take the derivatives and solve the non-linear system with fixed-point iterations.

Evidence Optimization: Practise in Python

To save the time, I skip the part with taking the derivatives and breaking system over old-new values. So, here there are iterative updates:

$$
\begin{aligned}
\mu^{new} &= \beta(\beta X^T X + A)^{-1} X^T \mathbf{t}, \\
\alpha_i^{new} &= \frac{1}{\mu_i^2}(1 - \Sigma_{ii}^{old} \alpha_i^{old}), \\
\beta^{new} &= \frac{1}{\|t - X\mu\|_2^2} \left(N - \text{trace}(I - \Sigma^{old} A^{old})\right), \\
\Sigma^{new} &= (\beta^{new} X^T X + A^{new})^{-1}.
\end{aligned}
$$

We want to prune feature i where $\alpha_i \to \infty$ Let's open the notebook in colab and implement this [click].

Ref.

- Tipping, M. E. (2000). The relevance vector machine. In Advances in neural information processing systems (pp. 652-658).
- Bishop, C. M., Tipping, M. (2013). Variational relevance vector machines. arXiv preprint arXiv:1301.3838.

All seminar materials and solutions are here, click.