CS281: Advanced ML **September 13, 2017**

Lecture 4: Linear Regression

Lecturer: Sasha Rush *Scribes: Kojin Oshiba, Michael Ge, Aditya Prasad, Scott Sussex*

4.1 Multivariate Normal (MVN)

The multivariate normal distribution of a D-dimensional random vector X is defined as:

$$\mathcal{N}(X;\mu,\Sigma) \sim (2\pi)^{-\frac{D}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right)$$

Note:

- $(2\pi)^{-\frac{D}{2}}|\Sigma|^{-\frac{1}{2}}$ and $-\frac{1}{2}$ are constants we can ignore in MLE and MAP calculations for μ.
- $(X-\mu)^T\Sigma^{-1}(X-\mu)$ is a quadratic term.

Given a set of X values, there are two types of inference on μ and Σ that we're interested in doing: MLE and MAP. We are also interested in doing predictions on future X values.

4.2 Maximum Likelihood of MVN

Let $\theta = (\mu, \Sigma)$, where Σ can be approximated as a diagonal/low rank matrix. If there are $x_1, \ldots, x_n$ observations, the MLE estimate of μ is

$$\begin{aligned}
\mu^* &= \arg\max_{\mu} -\sum_{i=0}^{n}\log\mathcal{N}(x_i;\mu,\Sigma) \\
&= \arg\max_{\mu} \quad \log(constant) - \sum_{i=0}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) \\
&= \arg\max_{\mu} -\sum_{i=0}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)
\end{aligned}$$

Let $L = \sum_{i=0}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$.

$$\begin{aligned}
\frac{dL}{d\mu} &= \Sigma_n\Sigma^{-1}(x_n-\mu) = 0 \\
&\Leftrightarrow \mu^*_{MLE} = \frac{\Sigma_{X_n}}{N}
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{dL}{d\Sigma} &= (exercise) = 0 \\
&\Leftrightarrow \Sigma^*_{MLE} = \frac{1}{N}\sum_{i=0}^{n}x_ix_i^T = \frac{1}{N}X^\top X
\end{aligned}$$

Exercise 4.1. Calculate $\frac{dL}{d\Sigma}$ The following might be helpful:

- $\frac{d}{dA} ln|A| = A^{-1}$
- $\frac{d}{dA} tr(BA) = B^T$
- $\text{tr}(ABC) = \text{tr}(BCA)$

Proof. (Solution also in Murphy page 102)
Define $\Delta = \Sigma^{-1}$ as the precision matrix. Let

$$\begin{aligned} L &= \frac{N}{2}\log(|\Delta|) - \frac{1}{2}\sum_{i=0}^{n}(x_i-\mu)^T\Delta(x_i-\mu) \\ &= \frac{N}{2}\log(|\Delta|) - \frac{1}{2}\sum_{i=0}^{n}\text{tr}((x_i-\mu)^T(x_i-\mu)\Delta \end{aligned}$$

Now

$$\frac{dL}{d\Gamma} = \frac{N}{2}\Delta^T - \frac{1}{2}X^TX$$

Δ is orthogonal so $\Delta^T = \Delta^{-1} = \Sigma$. This gives the result

$$\frac{dL}{d\Sigma} = \frac{N}{2}\Sigma - \frac{1}{2}X^TX$$

□

4.3 Linear-Gaussian Models

Let x be a vector of affine, noisy observations with a prior distribution:

$$x \sim N(m_0, S_0)$$

Let y be the outputs:

$$y|x \sim \mathcal{N}(Ax+b, \Sigma_y)$$

4.3.1 $p(x|y)$

We are interested in calculating the posterior distribution: $p(x|y)$.

$$\begin{aligned} p(x|y) &\propto p(x)p(y|x) \\ &= \frac{1}{2}\exp\begin{cases}(x-m_0)^\top S_0^{-1}(x-m_0) \\ +(y-(Ax+b))^\top\Sigma_y^{-1}(y-(Ax+b))\end{cases} \\ &= \frac{1}{2}\exp\begin{cases}x^\top S_0^{-1}x^{\star\star} - 2x^\top S_0^{-1}m_0{}^\star + \dots \\ \underline{+x^\top(A^\top\Sigma_y^{-1}A)x^{\star\star}} - \underline{2x^\top(A^\top\Sigma_y^{-1})y^\star} + \underline{2x^\top(A^\top\Sigma_y^{-1})b^\star} + \dots\end{cases} \end{aligned}$$

The terms containing x are underlined. Double-starred ($\star\star$) terms are quadratic in x, while single-starred ($\star$) terms are linear in x. The remaining terms are constants that are swallowed up by the proportionality. By Gaussian-Gaussian conjugacy, we know the resulting distribution should be Gaussian. To find the parameters, we'll modify $p(x|y)$ to fit the form of a Normal. This requires completing the square!

4.3.2 Completing the Square

$$ax^2 + bx + c \to a(x-h)^2 + k, h = \frac{-b}{2a}, k = c - \frac{b^2}{4a}$$

We ignore the k term since it too is swallowed up in the proportionality. In application to our problem, we group the quadratic and linear terms together to calculate our terms for completing the square.

- "a" is $S_N^{-1} = S_0^{-1} + A^\top \Sigma_y^{-1} A$
- "h" is $m_N = S_N \left[S_0^{-1} m_0 + A^\top \Sigma_y^{-1} (y - b) \right]$

In this more "intuitive" representation, we find that $p(x|y)$ has the form of $\mathcal{N}(m_N, S_N)$. Murphy also has a more explicit representation:

- $\Sigma_{x|y} = \Sigma_x^{-1} + A^\top \Sigma_y^{-1} A$
- $\mu_{x|y} = \Sigma_{x|y}[\Sigma_x^{-1}\mu_x + A^\top \Sigma_x^{-1}(y - b)]$

4.3.3 p(y)

We now calculate the normalizer term, $p(y)$. Now, x is fixed. y follows the linear model:

$$y = Ax + b + \epsilon$$

The result is that y follows a Normal distribution with the following form:

$$p(y) = \mathcal{N}(y; Am_0 + b, \Sigma_y + A\Sigma_x A^\top)$$

4.3.4 Prior (just for μ)

$$p(\mu) = \mathcal{N}(\mu | m_0, S_0)$$

where m_0, S_0 are prior mean, prior variance. $p(\mu)$ is defined Gaussian because Gaussian is the conjugate prior of itself. A prior is called a conjugate prior if it has the same distribution as the posterior distribution.

4.3.5 Posterior (just for μ)

$$p(\mu|X) \propto p(\mu)p(X|\mu) = \mathcal{N}(\mu|m_0, s_0)\mathcal{N}(X; \mu, \Sigma)$$

This is a special case of linear regression. Recall,

- "a" is $S_N^{-1} = S_0^{-1} + A^\top \Sigma_y^{-1} A$
- "h" is $m_N = S_N \left[S_0^{-1} m_0 + A^\top \Sigma_y^{-1} (y - b) \right]$

We let $b = 0$ and $A = I$. Then we obtain,

$$S_N^{-1} = S_0^{-1} + \Sigma^{-1}$$

$$m_N = S_N[S_0^{-1} m_0 + \Sigma^{-1} X]$$

Hence,

$$p(\mu|X) = \mathcal{N}(m; m_N, S_N)$$

4.3.6 Unknown Variance

Similar to μ, we can also define a conjugate prior on Σ, which is Inverse Wishart distribution. It is defined as:

$$IW(\Sigma|S,\nu) = \frac{1}{2}|\Sigma|^{-(\nu-(D+1)/2)} \exp\{\frac{1}{2}tr(S^{-1}\Sigma^{-1})\}$$

- distribution over positive semi definite Σ with two parameters S, ν.
- $S = \Sigma XX^T$ is a prior scatter matrix called the scale matrix. $\nu = n\mu$ is degrees of freedom where $\nu - (D+1)$ is the number of observations.

4.4 Linear Regression

In an undergraduate version of the class, we might define the problem as follows: We are given "fixed" set of inputs, $\{\boldsymbol{x}_i\}$. We want to "predict" the outputs.

Here, we define the problem as attempting to compute $p(y \mid \boldsymbol{x}, \boldsymbol{\theta})$. Consider the following example. We assume that our data is generated as follows:

$$y = \boldsymbol{w}^T\boldsymbol{x} + \text{noise}$$

Further, we assume that the noise (denoted by ϵ) is distributed as Gaussian with mean 0; that is:

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

Then, we have:

$$p(y \mid \boldsymbol{x}, \theta) = \mathcal{N}(y \mid \boldsymbol{w^T x}, \sigma^2)$$

Note that the bias term here is included as a dimension in $\boldsymbol{w}$, w_0.

4.4.1 Log Likelihood

Consider a data-set that looks like $\{(\boldsymbol{x_i}, y_i)\}_{i=1}^N$. The log-likelihood $\mathcal{L}(\boldsymbol{\theta})$ is given by:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \log p(\text{data} \mid \boldsymbol{\theta}) \\
&= \sum_{n=1}^{N} \log p(y_n \mid \boldsymbol{x_n}, \boldsymbol{\theta}) \\
&= \sum_{n=1}^{N} \log(\text{constant}) - \frac{1}{2\sigma^2}(y_n - \boldsymbol{w}^T\boldsymbol{x_n})^2
\end{aligned}$$

Note that data here refers to just the y_i's. The y_n's are called the target; the $\boldsymbol{w}$ represents the weights; and the x_n's are the observations. The term $(y_n - \boldsymbol{w}^T\boldsymbol{x_n})^2$ is essentially just the residual sum of squares.

4.4.2 Computing MLE

We want the argmax of the log-likelihood. We therefore have:

$$\begin{aligned}
\text{argmax}_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}) &= \text{argmax} - \sum_{n=1}^{N} \frac{1}{2\sigma^2}(y_n - \boldsymbol{w}^T\boldsymbol{x_n})^2 \\
&= \text{argmax}_{\boldsymbol{w}} - [\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}]^T [\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}] \\
&= \text{argmax}_{\boldsymbol{w}} \left[\boldsymbol{w}^T\boldsymbol{X^T}\boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{X^T}\boldsymbol{y} + \text{constant}\right]
\end{aligned}$$

There is an analytical solution to this, and we obtain it by simply computing the gradient and setting it to $\mathbf{0}$.

$$\partial_w \left[w^T \boldsymbol{X}^T \boldsymbol{X} w - 2 w^T \boldsymbol{X}^T \boldsymbol{y} \right] = 2\boldsymbol{X}^T \boldsymbol{X} w - 2\boldsymbol{X}^T \boldsymbol{y}$$

Setting this to $\mathbf{0}$, we obtain:

$$w_{MLE} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

As we will see in homework 1, $(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$ can be viewed as the projection of $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$.

4.5 Bayesian Linear Regression

In the Bayesian framework, we also introduce a probability distribution on the weights. Here, we choose:

$$p(w) = \mathcal{N}(w \mid \boldsymbol{m_0}, \boldsymbol{S_0})$$

Thus, we have:

$$p(\boldsymbol{y} \mid \boldsymbol{X}, w, \mu, \sigma^2) = \mathcal{N}(\boldsymbol{y} \mid \mu + \boldsymbol{X}^T w, \sigma^2 \boldsymbol{I})$$

We assume that $\mu = 0$.
The posterior then is of the form:

$$p(w \mid \ldots) \propto \mathcal{N}(w \mid \boldsymbol{m_0}, \boldsymbol{S_0}) \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{X}^T w, \sigma^2 \boldsymbol{I})$$

Applying the results obtained above with the linear Gaussian results, with:

$$\begin{aligned} b &= 0 \\ \boldsymbol{A} &= \boldsymbol{X}^T \\ \boldsymbol{\Sigma_y} &= \sigma^2 \boldsymbol{I} \end{aligned}$$

Thus, we have:

$$\begin{aligned} \boldsymbol{S_N}^{-1} &= \boldsymbol{S_0}^{-1} + \frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{X} \\ \boldsymbol{m_N} &= \boldsymbol{S_N} \left[\boldsymbol{S_0^{-1}} \boldsymbol{m_0} + \boldsymbol{X}^T \boldsymbol{y} \frac{1}{\sigma^2} \right] \end{aligned}$$

Now, we compute the posterior predictive:

$$p(y \mid x, y) = \int \mathcal{N}(y \mid w^T x, \sigma^2) \mathcal{N}(w \mid \boldsymbol{m_N}, \boldsymbol{S_N}) dw$$

Using the form for the marginal derived earlier, we have:

$$p(y \mid x, y) = \mathcal{N}(y \mid \boldsymbol{X}^T \boldsymbol{m_N}, \sigma^2 + \boldsymbol{X}^T \boldsymbol{S_0} \boldsymbol{X})$$

The variance term is particularly interesting because now the variance has dependence on the actual data; thus, the Bayesian method has thus produced a different result. The mean, however, is the same as the MAP estimate ($x^T \boldsymbol{m_N}$)

4.6 Non Linear Regression

All the examples done so far have been in linear space. To define an adaptive basis, we simply transform point x with the transormation of our choice:

$$x \rightarrow \phi(x)$$

Examples include:

- $\phi_1(x) = \sin(x)$
- $\phi_2(x) = \sin(\lambda x)$
- $\phi_3(x) = \max(0, x)$
- $\phi(x; w) = \max(0, w' \top x)$

The last example is the core of neural networks and deep learning where the weights are learned for each level of w.