CPSC 486/586: Probabilistic Machine Learning 01/25/2023

Lecture 4

Lecturer: Andre Wibisono *Scribe: Jorge Blanco*

1 Bayesian Statistics

In the bayesian approach to statistics we treat θ as an unknown parameter and the data as known We represent our uncertainty about the parameters after observing the data by calculating the **posterior distribution**. Let X denote the observed data then by Bayes' rule:

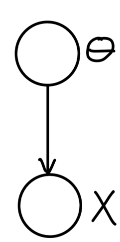
$$P(\theta|X) = \frac{P(\theta)P(X|\theta)}{P(X)}$$

where

1. $P(\theta|X)$ is the **posterior distribution**
2. $P(\theta)$ is our **prior** which represents our beliefs about the parameters before seeing the data.
3. $P(X|\theta)$ is called the **likelihood** and represents our beliefs about what data we expect to see for each setting of the parameters
4. $P(X)$ is the **marginal likelihood**. This is the same for all θ, so usually is ignored (under $\propto$) when maximizing the posterior with respect to θ. We obtain it by integrating over the parameter space (in contrast to the partition function, which we obtain by integrating over the x space).

1.1 Graphical representation:

We represent the dependence of X on θ as a directed graph:



2 Bayesian inference and linear regression

Let $\|x\|_A^2 = x^\top Ax$ and recall that the distribution $\mathcal{N}(\mu, \Sigma)$ for $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ has density:

$$f(x) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(\frac{-\|x-\mu\|_{\Sigma^{-1}}^2}{2}\right)$$

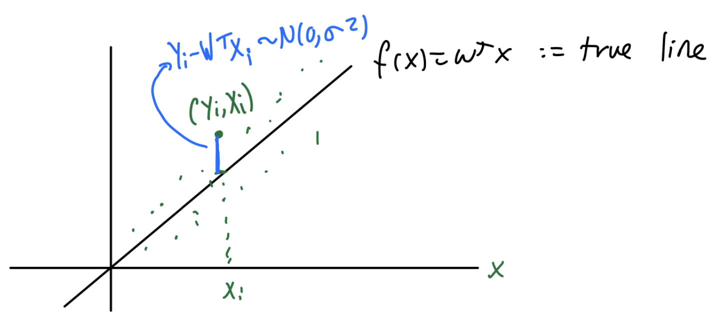
2.1 Model and Graphical Representation

The standard linear regression model is:

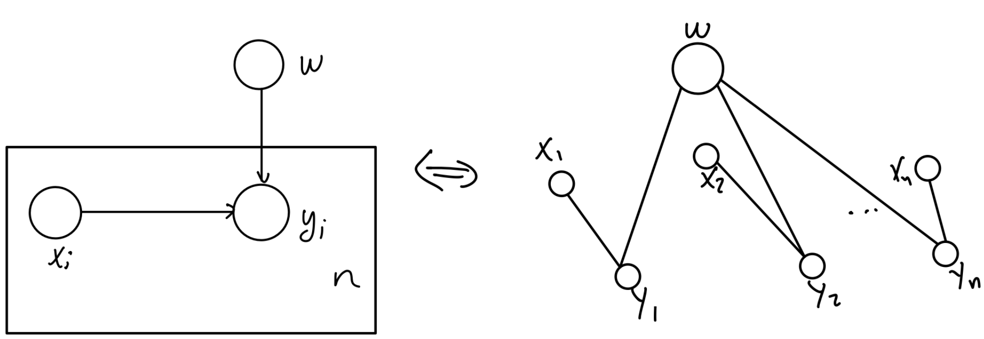
$$y_i = w^\top x_i + \epsilon_i$$

where $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$.

The picture associated with this model is:



and the graphical representation of the model is:



The box around the values x_i and y_i above mean that these dependencies repeat for $i \in [n]$. We can write the joint distribution of w, x, y as follows:

$$
\begin{aligned}
P(w,x,y) &= P(w)P(x_1,\ldots,x_n,y_1,\ldots,y_n|w) \\
&= P(w)\prod_{n=1}^{n} P(x_i,y_i|w)
\end{aligned}
$$

This equality holds because given w the pairs $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ are independent of each other. We can see this by the model description when we have w the y_i values depend on x_i and the corresponding independent error term. Similarly, by the model description, when given w

$$
y_i|w \sim \mathcal{N}(w^\top x_i, \sigma^2)
$$

Hence we can continue the computation above as follows:

$$
\begin{aligned}
P(w,x,y) &= P(w)P(x_1,\ldots,x_n,y_1,\ldots,y_n|w) \\
&= P(w)\prod_{n=1}^{n} P(x_i,y_i|w) \\
&= P(w)\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - w^\top x_i\right)^2\right) \\
&= \frac{P(w)}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - w^\top x_i\right)^2\right)
\end{aligned}
$$

Note that in the exponential term above we have the expression $\sum_{i=1}^{n}\left(y_i - w^\top x_i\right)^2$ which we should recognize as the objective function of the ordinary least squares problem.

By Bayes rule, the posterior distribution of w given (x,y) also has the same form, but now we only care about the dependence on w:

$$
\begin{aligned}
P(w \mid x,y) &\propto P(w,x,y) \\
&\propto P(w)\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - w^\top x_i\right)^2\right). \qquad (1)
\end{aligned}
$$

Note that in (1) above, we have the prior term $P(w)$ let us consider some possible assignments to it:

(1) No prior

Suppose $P(w) = 1$, then the maximum likelihood estimator yields:

$$\begin{aligned}\arg\max_w \; P(x, y|w) &= \arg\max_w \; \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - w^T x_i\right)^2\right) \\ &= \arg\min_w \; \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - w^T x_i\right)^2 \\ &= \text{linear regression}\end{aligned}$$

(2) Gaussian prior

Suppose we use Gaussian Prior:

$$P(w) = \mathcal{N}(0, \lambda I)$$

for some $\lambda > 0$. Now we have that:

$$\begin{aligned}P(w|x, y) &\propto P(w)P(x, y|w) \\ &\propto \exp\left(-\frac{\|w\|^2}{2\lambda} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - w^\top x_i)\right)\end{aligned}$$

Notice that this is a Gaussian distribution as the terms inside the exponential can be written as quadratic form $(w - w^*)^\top A(w - w^*)$ like we did in the previous lecture. Suppose we want to approximate this.

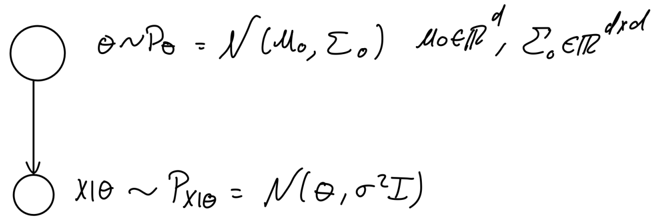
MAP - Maximum a Posteriori

A simple approximation is via the mode, i.e. the point which maximizes the posterior. This is called the MAP (Maximum a posteriori) estimator:

$$\begin{aligned}w_{map} &= \arg\max_w \; P(w|x, y) \\ &= \arg\max_w \; \exp\left(-\frac{\|w\|^2}{2\lambda} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - w^t x_i)\right) \\ &= \arg\max_w \; \exp\left(-\frac{1}{\sigma^2}\left(\frac{\sigma^2\|w\|^2}{2\lambda} + \frac{1}{2}\sum_{i=1}^{n}(y_i - w^t x_i)\right)\right) \\ &= \arg\min_w \left(\frac{\sigma^2\|w\|^2}{2\lambda} + \frac{1}{2}\sum_{i=1}^{n}(y_i - w^t x_i)\right)\end{aligned}$$

the above is equivalent to linear regression with ℓ_2 regularization, also known as ridge regression (with regularization parameter σ^2/λ). Now let us consider a more general example.

Example 1. *Let* $\theta \sim P_\theta = \mathcal{N}(\mu_0, \Sigma_0)$ *and* $x|\theta \sim P_{x|\theta} = \mathcal{N}(\theta, \sigma^2 I)$ *we call this a Gaussian model.*



We find the posterior of the Gaussian model above note that:

$$
\begin{aligned}
P(\theta|X) &\propto P(\theta)P(X|\theta) \\
&\propto \exp\left(-\frac{\|\theta - \mu_0\|^2_{\Sigma_0^{-1}}}{2} - \frac{\|x - \theta\|^2}{2\sigma^2}\right) \\
&\propto \exp\left[-\frac{\langle \theta, \Sigma_0^{-1}\theta \rangle}{2} + \langle \theta, \Sigma_0^{-1}\mu_0 \rangle - \frac{\langle \theta, \theta \rangle}{2\sigma^2} + \frac{\langle \theta, x \rangle}{\sigma^2}\right] \qquad \text{(note: dropped terms not depending on } \theta\text{)} \\
&\propto \exp\left[-\frac{1}{2}\langle \theta, \left(\Sigma_0^{-1} + \frac{1}{\sigma^2} I\right)\theta \rangle + \langle \theta, \Sigma_0^{-1}\mu_0 + \frac{1}{\sigma^2}x \rangle\right] \\
&\propto \exp\left[-\frac{\|\theta - \mu_1\|^2_{\Sigma_1^{-1}}}{2}\right]
\end{aligned}
$$

where

$$\Sigma_1^{-1} = \Sigma_0^{-1} + \frac{1}{\sigma^2} I$$

and

$$\mu_1 = \Sigma_1 \left(\Sigma_0^{-1}\mu_0 + \frac{1}{\sigma^2} x\right).$$

Therefore:

$$P(\theta|X) = N(\mu_1, \Sigma_1)$$

Moreover the quantities above satisfy

$$\mu_1 = \left(\frac{\frac{1}{\Sigma_0}}{\frac{1}{\Sigma_0} + \frac{1}{\sigma^2}}\right)\mu_0 + \left(\frac{\frac{1}{\sigma^2}}{\frac{1}{\Sigma_0} + \frac{1}{\sigma^2}}\right)x$$

Consider the following observations/consequences:

1. μ_1 is a convex combination of μ_0 and x

2. if $\sigma^2 \to \infty$ we learn nothing from the data as just get $\mu_1 = \mu_0$

3. if $\sigma^2 \to 0$ is this is the case our posterior mean is just X which makes sense because as $\sigma \to 0$ we have $P(X|\theta) \to \delta_\theta$ (a point mass at θ)

Gaussian Model with multiple observations

Consider the following generalization of the model above (with more observations):

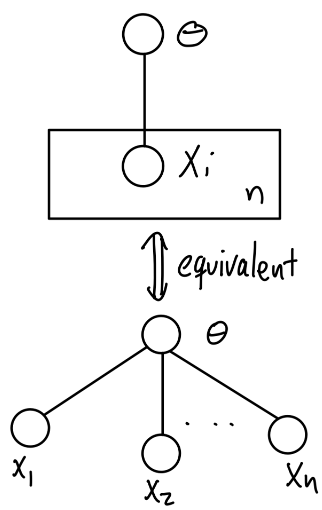
$$\theta \sim P_0 = \mathcal{N}(\mu_0, \Sigma_0)$$

for each $i \in [n]$:

$$x_i|\theta \sim \mathcal{N}(\theta, \sigma^2 I)$$

independently.

Graphically we can express the model as:



Using the model's specifications we can write:

$$P(\theta, x_1, \ldots, x_n) = P(\theta) \prod_{i=1}^{n} P(x_i|\theta)$$

hence

$$P(\theta|x_1, \ldots, x_n) \propto_\theta P(\theta) \prod_{i=1}^{n} P(x_i|\theta)$$

note that we can also write:

$$P(\theta|x_1, \ldots, x_{n-1}, x_n) \propto P(\theta|x_1, \ldots, x_{n-1}) P(x_n|\theta)$$

A similar computation as the one above yields:

$$P(\theta|x_1,\ldots,x_n) = \mathcal{N}(\mu_n, \Sigma_n)$$

for

$$\Sigma_n^{-1} = \Sigma_0^{-1} + \frac{n}{\sigma^2} I$$

and

$$\Sigma_n^{-1}\mu_n = \Sigma_0^{-1}\mu_0 + \frac{1}{\sigma^2}(x_1 + \ldots + x_n)$$

these parameters have some interesting asymptotic properties: as $n \to \infty$ we have that:

1. $\Sigma_n^{-1} \to \infty$ (equivalently, $\Sigma_n \to 0$.
2. $\Sigma_n^{-1}\mu_n = \Sigma_0^{-1}\mu_0 + \frac{n}{\sigma^2}\overline{x}_n$ where $\overline{x}_n = \frac{1}{n}(x_1 + \cdots + x_n)$ is the sample mean. Now if $x_1, \ldots, x_n$ are generated from some distribution, the sample mean converges to the true mean, $\overline{x}_n \to \mathbb{E}[x]$ as $n \to \infty$. Since $\Sigma_n^{-1} = O(n)$, from the above you can show that $\mu_n \to \mathbb{E}[x]$ as $n \to \infty$.

Definition 1 (Exponential Families)**.** *We say a distribution is in the exponential family if its density is of the form*

$$P_\theta(x) = h(x)\exp(\langle \theta, T(X)\rangle - A(\theta))$$

for some base measure h on $\mathbb{R}^d$, sufficient statistic $T : \mathbb{R}^d \to \mathbb{R}^m$, canonical parameter $\theta \in \Theta \subset \mathbb{R}^m$, where $A : \Theta \to \mathbb{R}$ is the log partition function:

$$A(\theta) = \log\left(\int_{\mathbb{R}^d} e^{\langle \theta, T(x)\rangle} h(x)\, dx\right)$$

and the domain is $\Theta = \{\theta \in \mathbb{R}^m : A(\theta) < \infty\}$.

Example 2 (The normal is in exponential family)**.** *The multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ is in the exponential family:*

Proof. We can write the normal density $\mathcal{N}(\mu, \Sigma)$ as

$$\begin{aligned}
p(x) &= \frac{1}{\sqrt{\det(2\pi\Sigma)}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)} \\
&= \exp\left(-\frac{1}{2}x^\top\Sigma^{-1}x + x^\top\Sigma^{-1}\mu - \frac{1}{2}\mu^\top\Sigma^{-1}\mu - \frac{1}{2}\log\det(2\pi\Sigma)\right) \\
&= \exp\left(-\frac{1}{2}\langle xx^\top, \Sigma^{-1}\rangle_F + \langle x, \Sigma^{-1}\mu\rangle - \frac{1}{2}\mu^\top\Sigma^{-1}\mu - \frac{1}{2}\log\det(2\pi\Sigma)\right) \\
&= \exp\left(\langle T(x), \theta\rangle - A(\theta)\right)
\end{aligned}$$

where the sufficient statistic is given by

$$T(x) = \left(x, -\frac{1}{2}xx^\top\right) \in \mathbb{R}^{d+d^2}$$

and the parameter $\theta \in \mathbb{R}^{d+d^2}$ is given by

$$\theta = (\Sigma^{-1}\mu, \Sigma^{-1})$$

so that the inner product is given by

$$\langle T(x), \theta\rangle = \langle x, \Sigma^{-1}\mu\rangle + \left\langle -\frac{1}{2}xx^\top, \Sigma^{-1}\right\rangle_F.$$

(In the above, $\langle A, B\rangle_F = \mathsf{Tr}(AB^\top)$ is the Frobenius inner product of two matrices, which is equivalent to the ℓ_2-inner product of the "vectorized" version of the matrices: $\langle A, B\rangle_F = \sum_{i,j=1}^d A_{ij}B_{ij}$.) The log-partition function is given by

$$A(\theta) = \frac{1}{2}\mu^\top\Sigma^{-1}\mu + \frac{1}{2}\log\det(2\pi\Sigma).$$

From the above, we see why the inverse covariance Σ^{-1} and the normalized mean $\Sigma^{-1}\mu$ are important quantities in the calculations, because they are the canonical parameters of the Gaussian distribution as an exponential family distribution.

$\square$