Bayesian Approach to Machine Learning (4)

Rohit Budhiraja

Machine Learning for Wireless Communications (EE798L)

Feb. 2, 2024

Recap of last lecture and today's agenda

- Recap of last class
  - Finished discussing Bayesian framework for coin toss example
- Today's agenda
  - Bayesian framework for Olympic data

Bayesian treatment for coin tossing game (recap)

- Modelled data using binomial distribution with likelihood

$$P\left(Y=y|r,N\right)=\left(\begin{array}{c}N\\ y\end{array}\right)r^{y}(1-r)^{N-y}$$

- Calculated winning probability with a ML (point) estimate of r i.e., $P\left(Y_{new}\leq 6|\hat{r}_{ML}\right)$
- Considered r as a random variable, and captured its uncertainty
- By defining random variable Y_N as number of heads obtained in N tosses, we calculated $p\left(r|y_N\right)$

$$p\left(r|y_N\right)=\frac{P\left(y_N|r\right)p\left(r\right)}{P\left(y_N\right)}$$

- Re-calculated winning probability with MAP estimate of r i.e., $P\left(Y_{new}\leq 6|\hat{r}_{MAP}\right)$
- Computed expected winning probability using all of posterior information as well

$$\int_{r=0}^{r=1}P\left(Y_{new}\leq 6|r\right)p\left(r|y_N\right)dr=\mathbb{E}_{p(r|y_N)}\{P\left(Y_{new}\leq 6|r\right)\}$$

- Computed marginal likelihood, and briefly informed that it captures model information

Maximum likelihood approach for Olympic data (recap)

- Recall our Olympic data model: $t_n = \mathbf{w}^T \mathbf{x}_n + \epsilon_n$ such that

$$
\begin{aligned}
t_1 &= w_0 + x_1 w_1 + \epsilon_1 \\
\vdots &= \vdots \\
t_N &= w_0 + x_N w_1 + \epsilon_N \\
\mathbf{t} &= \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}
\end{aligned}
$$

- Joint Gaussian likelihood of N data points

$$
L = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^T \mathbf{x}_n, \sigma^2) = \mathcal{N}\left(\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}_N\right)
$$

- Calculated point ML estimate of $\mathbf{w}$ by maximizing logarithm likelihood which is $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$
- Predicted $t_{new} = \widehat{\mathbf{w}}^T \mathbf{x}_{new}$, and calculated predictive variance $\sigma^2_{new} = var\{t_{new}\}$
- Bayesian approach: calculate posterior distribution of $\mathbf{w}$, and use it to calculate t_{new}

Outline of Bayesian treatment of Olympic data

- We will use kth order polynomial model to model Olympic data

$$t_n = w_0 + w_1 x_n + w_2 x_n^2 + \ldots + w_K x_n^K + \epsilon_n = \mathbf{w}^T \mathbf{x}_n + \epsilon_n$$

  where $\mathbf{w} = [w_0, \ldots, w_k]^T$ and $\mathbf{x}_n = [1, x_n, x_n^2, \ldots, x_n^K]^T$, $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$
- We can also write

$$\mathbf{t} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$$

  where $\mathbf{t} = [t_1, \ldots, t_N]^T$, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T$ and $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^T$
- Bayesian treatment enables us to incorporate insights about $\mathbf{w}$ using a prior
  - Will define prior over $\mathbf{w}$ using set of parameters Δ
- We assume that true value of σ^2 is known – simplify math
- Characterize randomness of $\mathbf{w}$ by calculating posterior distribution using Bayes rule

$$p\left(\mathbf{w}|\mathbf{t}\right) = \frac{p\left(\mathbf{t}|\mathbf{w}\right) p\left(\mathbf{w}\right)}{p\left(\mathbf{t}\right)}$$

- Posterior of $\mathbf{w}$ with dependence on parameter and data made explicit

$$p\left(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta\right) = \frac{p\left(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2, \Delta\right) p\left(\mathbf{w}|\Delta\right)}{p\left(\mathbf{t}|\mathbf{X}, \sigma^2, \Delta\right)} = \frac{p\left(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2\right) p\left(\mathbf{w}|\Delta\right)}{p\left(\mathbf{t}|\mathbf{X}, \sigma^2, \Delta\right)}$$

Olympic data – choice of prior

- For our model $\mathbf{t} = \mathbf{X}\mathbf{w} + \epsilon$, likelihood is Gaussian $p\left(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2\right) = \mathcal{N}\left(\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}_N\right)$
- Use a Gaussian prior $p\left(\mathbf{w}|\Delta\right)$, which is conjugate to Gaussian likelihood,

$$p\left(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) = \mathcal{N}\left(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right),$$

- We will choose parameters $\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$ later. Also, we will not always explicitly condition on $\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$
- For example, instead of $p\left(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta\right) = p\left(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)$ we will use $p\left(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2\right)$

Olympic data – Posterior calculation (1)

- We will use the fact that posterior will be Gaussian – allows us to ignore the marginal likelihood

$$
\begin{aligned}
p\left(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2\right) &\propto p\left(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2\right) p\left(\mathbf{w}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0\right) \\
&= \frac{1}{(2\pi)^{N/2}\left|\sigma^2\mathbf{I}\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{t}-\mathbf{X}\mathbf{w}\right)^T\left(\sigma^2\mathbf{I}\right)^{-1}\left(\mathbf{t}-\mathbf{X}\mathbf{w}\right)\right) \\
&\times \frac{1}{(2\pi)^{N/2}\left|\boldsymbol{\Sigma}_0\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{w}-\boldsymbol{\mu}_0\right)^T\boldsymbol{\Sigma}_0^{-1}\left(\mathbf{w}-\boldsymbol{\mu}_0\right)\right) \\
&\overset{(a)}{\propto} \exp\left(-\frac{1}{2\sigma^2}\left(\mathbf{t}-\mathbf{X}\mathbf{w}\right)^T\left(\mathbf{t}-\mathbf{X}\mathbf{w}\right)\right) \times \exp\left(-\frac{1}{2}\left(\mathbf{w}-\boldsymbol{\mu}_0\right)^T\boldsymbol{\Sigma}_0^{-1}\left(\mathbf{w}-\boldsymbol{\mu}_0\right)\right) \\
&= \exp\left\{-\frac{1}{2}\left(-\frac{1}{\sigma^2}\left(\mathbf{t}-\mathbf{X}\mathbf{w}\right)^T\left(\mathbf{t}-\mathbf{X}\mathbf{w}\right)+\left(\mathbf{w}-\boldsymbol{\mu}_0\right)^T\boldsymbol{\Sigma}_0^{-1}\left(\mathbf{w}-\boldsymbol{\mu}_0\right)\right)\right\}.
\end{aligned}
$$

- Proportionality (a) is because we ignore the terms which are independent of $\mathbf{w}$
- Multiplying the terms in bracket, and once again removing any that don't involve $\mathbf{w}$ gives

$$
p\left(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2\right) \propto \exp\left\{-\frac{1}{2}\left(-\frac{2}{\sigma^2}\mathbf{t}^T\mathbf{X}\mathbf{w}+\frac{1}{\sigma^2}\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}+\mathbf{w}^T\boldsymbol{\Sigma}_0^{-1}\mathbf{w}-2\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\mathbf{w}\right)\right\}
$$

Olympic data – Posterior calculation (2)

- We have

$$p\left(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2\right) \propto \exp\left\{-\frac{1}{2}\left(-\frac{2}{\sigma^2}\mathbf{t}^T\mathbf{X}\mathbf{w}+\frac{1}{\sigma^2}\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}+\mathbf{w}^T\boldsymbol{\Sigma}_0^{-1}\mathbf{w}-2\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\mathbf{w}\right)\right\} \tag{1}$$

- We take a generic multivariate Gaussian pdf and rearrange it to make it look like (1)

$$\begin{aligned} p\left(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2\right) &= \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{w}},\boldsymbol{\Sigma}_{\mathbf{w}}\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\mathbf{w}-\boldsymbol{\mu}_{\mathbf{w}}\right)^T\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\left(\mathbf{w}-\boldsymbol{\mu}_{\mathbf{w}}\right)\right) \\ &\propto \exp\left\{-\frac{1}{2}\left(\mathbf{w}^T\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{w}-2\boldsymbol{\mu}_{\mathbf{w}}^T\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{w}\right)\right\} \end{aligned} \tag{2}$$

- Linear and quadratic term in $\mathbf{w}$ in (1) must be equal to those in (2). Start with quadratic

$$\begin{aligned} \mathbf{w}^T\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{w} &= \frac{1}{\sigma^2}\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}+\mathbf{w}^T\boldsymbol{\Sigma}_0^{-1}\mathbf{w}=\mathbf{w}^T\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}+\boldsymbol{\Sigma}_0^{-1}\right)\mathbf{w} \\ \boldsymbol{\Sigma}_{\mathbf{w}} &= \left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}+\boldsymbol{\Sigma}_0^{-1}\right)^{-1} \end{aligned}$$

Olympic data – Posterior calculation (3)

- Similarly, equating linear terms from posterior we can get an expression for $\boldsymbol{\mu}_{\mathbf{w}}$:

$$
\begin{aligned}
-2\boldsymbol{\mu}_{\mathbf{w}}^T\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{w} &= -\frac{2}{\sigma^2}\mathbf{t}^T\mathbf{X}\mathbf{w} - 2\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\mathbf{w} \\
\boldsymbol{\mu}_{\mathbf{w}}^T\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{w} &= \frac{1}{\sigma^2}\mathbf{t}^T\mathbf{X}\mathbf{w} + \boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\mathbf{w} \\
\boldsymbol{\mu}_{\mathbf{w}}^T\boldsymbol{\Sigma}_{\mathbf{w}}^{-1} &= \frac{1}{\sigma^2}\mathbf{t}^T\mathbf{X} + \boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1} \\
\boldsymbol{\mu}_{\mathbf{w}}^T\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\boldsymbol{\Sigma}_{\mathbf{w}} &= \left(\frac{1}{\sigma^2}\mathbf{t}^T\mathbf{X} + \boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\right)\boldsymbol{\Sigma}_{\mathbf{w}} \\
\boldsymbol{\mu}_{\mathbf{w}}^T &= \left(\frac{1}{\sigma^2}\mathbf{t}^T\mathbf{X} + \boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\right)\boldsymbol{\Sigma}_{\mathbf{w}}
\end{aligned}
$$

- Using the fact $\boldsymbol{\Sigma}_{\mathbf{w}}^T = \boldsymbol{\Sigma}_{\mathbf{w}}$, we have

$$
\boldsymbol{\mu}_{\mathbf{w}} = \boldsymbol{\Sigma}_{\mathbf{w}}\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{t} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right),
$$

Olympic data – Posterior calculation (4)

- Posterior is therefore

$$p\left(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2\right) = \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}}\right)$$

with

$$\boldsymbol{\Sigma}_{\mathbf{w}} = \left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}$$

and

$$\boldsymbol{\mu}_{\mathbf{w}} = \boldsymbol{\Sigma}_{\mathbf{w}}\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{t} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)$$

- Mean (mode) $\hat{\mathbf{w}} = \boldsymbol{\mu}_{\mathbf{w}}$ is (MAP) estimate. If prior has zero mean $\boldsymbol{\mu}_0 = \mathbf{0}$ then MAP estimate is

$$\boldsymbol{\mu}_{\mathbf{w}} = \boldsymbol{\Sigma}_{\mathbf{w}}\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{t}\right) = \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}\mathbf{X}^T\mathbf{t}$$

  - Similar to regularized least squares estimate
- Given a new observation $\mathbf{x}_{new}$, predict t_{new} using point estimate
- Treat t_{new} as a random variable by casting $t_{new} = \mathbf{w}^T\mathbf{x}_{new} + \epsilon_{new}$

$$p\left(t_{new}|\mathbf{x}_{new}, \hat{\mathbf{w}}, \sigma^2\right) = \mathcal{N}\left(\hat{\mathbf{w}}^T\mathbf{x}_{new}, \sigma^2\right)$$

- Using MAP estimate, $t_{new} = \hat{\mathbf{w}}^T\mathbf{x}_{new}$, and has variance σ^2

Predictive distribution calculation (summary)

- Predict t_{new} by capturing complete randomness in $\mathbf{w}$ using its posterior distribution

$$\int p\left(t_{new}|\mathbf{x}_{new}, \mathbf{w}, \sigma^2\right) p\left(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2\right) d\mathbf{w} = p\left(t_{new}|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}, \sigma^2\right)$$

- $p\left(t_{new}|\cdot\right)$, also called predictive distribution, and is not conditioned on $\mathbf{w}$
- Recall $t_{new} = \mathbf{w}^T \mathbf{x}_{new} + \epsilon_{new}$ with

$$p\left(t_{new}|\mathbf{x}_{new}, \mathbf{w}, \sigma^2\right) = \mathcal{N}\left(\mathbf{x}_{new}^T \mathbf{w}, \sigma^2\right) \text{ with } p\left(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2\right) = \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}}\right)$$

Theorem

If marginal and conditional distributions of a generic Gaussian vector $\mathbf{x}$ are

$$p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \text{ and } p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}).$$

Here $\boldsymbol{\Lambda}$ and $\mathbf{L}$ are precision (inverse of covariance) matrices. Marginal distribution of $\mathbf{y}$ is

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T)$$

- Assume $\mathbf{w} = \mathbf{x}$ and $t_{new} = y$ and with $\boldsymbol{\mu} = \boldsymbol{\mu}_{\mathbf{w}}$, $\boldsymbol{\Lambda}^{-1} = \boldsymbol{\Sigma}_{\mathbf{w}}$, $b = 0$, $\mathbf{L}^{-1} = \sigma^2$, $\mathbf{A} = \mathbf{x}_{new}^T$, we have

$$p\left(t_{new}|\mathbf{x}_{new}\mathbf{X}, \mathbf{t}, \sigma^2\right) = \int p\left(t_{new}|\mathbf{x}_{new}, \mathbf{w}, \sigma^2\right) p\left(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2\right) d\mathbf{w} = \mathcal{N}\left(\mathbf{x}_{new}^T \boldsymbol{\mu}_{\mathbf{w}}, \sigma^2 + \mathbf{x}_{new}^T \boldsymbol{\Sigma}_{\mathbf{w}} \mathbf{x}_{new}\right)$$

- $t_{new} = \mathbf{x}_{new}^T \boldsymbol{\mu}_{\mathbf{w}}$ with variance $\sigma^2 + \mathbf{x}_{new}^T \boldsymbol{\Sigma}_{\mathbf{w}} \mathbf{x}_{new}$

Predicting using MAP estimate of w and its complete posterior

- Given a new observation $\mathbf{x}_{new}$, we want to predict t_{new}
- Treat t_{new} as a random variable by casting $t_{new} = \mathbf{w}^T\mathbf{x}_{new} + \epsilon_{new}$

$$p\left(t_{new}|\mathbf{x}_{new}, \mathbf{w}, \sigma^2\right) = \mathcal{N}\left(\mathbf{w}^T\mathbf{x}_{new}, \sigma^2\right) \tag{3}$$

- With MAP estimate $\hat{\mathbf{w}}$, Eq. (3) using $\hat{\mathbf{w}}$, is $p\left(t_{new}|\mathbf{x}_{new}, \hat{\mathbf{w}}, \sigma^2\right) = \mathcal{N}\left(\hat{\mathbf{w}}^T\mathbf{x}_{new}, \sigma^2\right)$
  - $t_{new} = \hat{\mathbf{w}}^T\mathbf{x}_{new} = \boldsymbol{\mu}_{\mathbf{w}}^T\mathbf{x}_{new}$ with variance σ^2
- With complete distribution of $\mathbf{w}$, we calculated predictive distribution

$$\int p\left(t_{new}|\mathbf{x}_{new}, \mathbf{w}, \sigma^2\right) p\left(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2\right) d\mathbf{w} = p\left(t_{new}|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}, \sigma^2\right)$$

- We showed $p\left(t_{new}|\mathbf{x}_{new}\mathbf{X}, \mathbf{t}, \sigma^2\right) = \mathcal{N}\left(\mathbf{x}_{new}^T\boldsymbol{\mu}_{\mathbf{w}}, \sigma^2 + \mathbf{x}_{new}^T\boldsymbol{\Sigma}_{\mathbf{w}}\mathbf{x}_{new}\right)$
  - Implies $t_{new} = \mathbf{x}_{new}^T\boldsymbol{\mu}_{\mathbf{w}}$ with variance $\sigma^2 + \mathbf{x}_{new}^T\boldsymbol{\Sigma}_{\mathbf{w}}\mathbf{x}_{new}$