Probabilistic Linear Regression

Piyush Rai
IIT Kanpur

Probabilistic Machine Learning (CS772A)

Jan 13, 2016

Linear Regression: A Probabilistic View

- Given: N training examples $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$, features: $\boldsymbol{x}_n \in \mathbb{R}^D$, response $y_n \in \mathbb{R}$
- $\mathbf{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_N]^\top$: $N \times D$ feat. matrix, $\mathbf{Y} = [y_1 \dots y_N]^\top$: $N \times 1$ resp. vector
- Probabilistic view: responses are generated via a probabilistic model
- Assume a "noisy" linear model with regression weight vector $\boldsymbol{w} \in \mathbb{R}^D$:

$$\boxed{y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n}$$

- Gaussian noise: $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$, β: precision (inverse variance) of Gaussian
- Thus each response y_n also has a Gaussian distribution

$$\boxed{y_n \sim \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})}$$

- Goal: Learn regression weight vector $\boldsymbol{w}$ to predict y_* for a new $\boldsymbol{x}_*$

Linear Regression: A Probabilistic View

- For Gaussian response y_n

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \sqrt{\frac{\beta}{2\pi}} \exp\left\{-\frac{\beta}{2}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2\right\}$$

- Thus the likelihood (assuming i.i.d. responses) or *probability* of data:

$$p(\mathbf{Y}|\mathbf{X}, \boldsymbol{w}) = \prod_{n=1}^{N} p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \exp\left\{-\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2\right\}$$

- Note: $\boldsymbol{x}_n$ (features) assumed given/fixed. Only modeling the response y_n
- **Log-likelihood** (ignoring constants w.r.t. $\boldsymbol{w}$)

$$\boxed{\log p(\mathbf{Y}|\mathbf{X}, \boldsymbol{w}) \propto -\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2}$$

- Note that the log-likelihood is nothing but a (weighted) sum of (negative) squared errors on training data: high log-lik $\Rightarrow$ low sum of squared errors

Maximum Likelihood Estimation (MLE)

- MLE: Find the $\boldsymbol{w}$ that maximizes the (log) likelihood $\log p(\mathbf{Y}|\mathbf{X}, \boldsymbol{w})$

$$\arg\max_{\boldsymbol{w}} \log p(\mathbf{Y}|\mathbf{X}, \boldsymbol{w}) = \arg\min_{\boldsymbol{w}} - \log p(\mathbf{Y}|\mathbf{X}, \boldsymbol{w}) = \arg\min_{\boldsymbol{w}} \frac{\beta}{2} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2$$

- Same objective as the classic **ordinary least squares** (OLS) regression
  - Basically, maximizing log-lik = minimizing the sum of squared errors
- Taking derivative w.r.t. $\boldsymbol{w}$ and setting to zero, we get

$$\boxed{\boldsymbol{w}_{MLE} = (\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top)^{-1} \sum_{n=1}^{N} y_n \boldsymbol{x}_n = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}}$$

- Same solution as the solution of the OLS regression problem. Some issues:
  - $\mathbf{X}^\top \mathbf{X}$ may be ill-conditioned (not invertible)
  - "Uncontrolled" $\boldsymbol{w}$ can lead to overfitting (thus need regularization)
- A solution: Put a prior distribution on $\boldsymbol{w}$ (to impose "smoothness" and control $\boldsymbol{w}$) and do MAP estimation (MAP estimation = "regularized" MLE)

Prior Distribution on Weights

- Assume zero-mean spherical Gaussian prior on weights $\boldsymbol{w} = [w_1 \ w_2 \ldots \ w_D]$

$$p(\boldsymbol{w}) = \mathcal{N}(0, \lambda^{-1}\mathbf{I}_D) = \left(\frac{\lambda}{2\pi}\right)^{D/2} \exp(-\frac{\lambda}{2}\boldsymbol{w}^\top \boldsymbol{w}) = \left(\frac{\lambda}{2\pi}\right)^{D/2} \exp(-\frac{\lambda}{2}||\boldsymbol{w}||^2)$$

  λ is **precision** (inverse variance) of the Gaussian and $||\boldsymbol{w}||^2 = \sum_{d=1}^{D} w_d^2$

- Note: We can also write the prior as a product of D univariate Gaussians

$$p(\boldsymbol{w}) = \prod_{d=1}^{D} p(w_d) = \prod_{d=1}^{D} \mathcal{N}(0, \lambda^{-1}) = \prod_{d=1}^{D} \sqrt{\frac{\lambda}{2\pi}} \exp(-\frac{\lambda}{2} w_d^2) = \left(\frac{\lambda}{2\pi}\right)^{D/2} \exp(-\frac{\lambda}{2}\sum_{d=1}^{D} w_d^2)$$

- Gaussian prior encourages a "small" $\boldsymbol{w}$ by shrinking each component w_d towards zero (Gaussian's mean). Precision λ controls the extent of shrinkage
- This corresponds to imposing a regularizer on $\boldsymbol{w}$. We will soon see (or you might already have guessed) that the Gaussian prior results in a squared norm (ℓ_2) regularizer, and λ controls the strength of regularization
- Note: Different types of priors result in different types of regularizers (e.g., a Laplace prior on $\boldsymbol{w}$: $p(\boldsymbol{w}) \propto \exp(-|\boldsymbol{w}|)$ will result in an ℓ_1 regularizer on $\boldsymbol{w}$)

MAP Estimation

- The posterior distribution on $\boldsymbol{w}$: $p(\boldsymbol{w}|\mathbf{X},\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X},\boldsymbol{w})p(\boldsymbol{w})$
- The (log) posterior: $\log p(\boldsymbol{w}|\mathbf{X},\mathbf{Y}) = \log p(\mathbf{Y}|\mathbf{X},\boldsymbol{w}) + \log p(\boldsymbol{w})$. Thus,

$$\log p(\boldsymbol{w}|\mathbf{X},\mathbf{Y}) \quad \propto \quad -\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 - \frac{\lambda}{2}\boldsymbol{w}^\top \boldsymbol{w} \quad \text{(ignoring constants w.r.t } \boldsymbol{w}\text{)}$$

- MAP Estimation: Maximize the (log) posterior w.r.t. $\boldsymbol{w}$

$$\arg\max_{\boldsymbol{w}} \log p(\boldsymbol{w}|\mathbf{X},\mathbf{Y}) = \arg\min_{\boldsymbol{w}} -\log p(\boldsymbol{w}|\mathbf{X},\mathbf{Y}) = \arg\min_{\boldsymbol{w}} \underbrace{\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2}_{\text{fit to the training data}} + \underbrace{\frac{\lambda}{2}\boldsymbol{w}^\top \boldsymbol{w}}_{\text{keep } \boldsymbol{w} \text{ “simple”}}$$

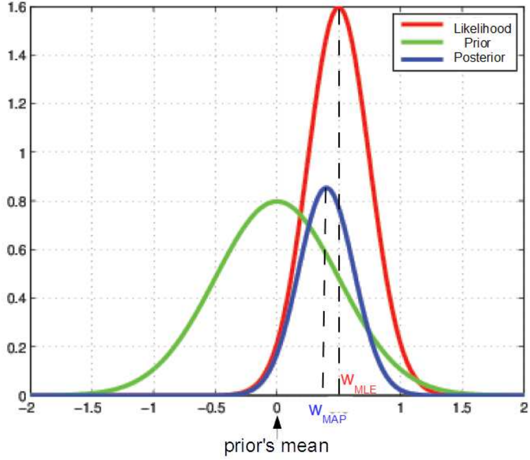
- Thus MAP estimation finds a $\boldsymbol{w}$ by trying to balance between the likelihood (fit to the training data) vs the prior (model's simplicity)
- Setting derivative w.r.t. $\boldsymbol{w}$ to zero yields

$$\boxed{\boldsymbol{w}_{MAP} = (\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top + \frac{\lambda}{\beta}\mathbf{I}_D)^{-1}\sum_{n=1}^{N} y_n\boldsymbol{x}_n = (\mathbf{X}^\top\mathbf{X} + \frac{\lambda}{\beta}\mathbf{I}_D)^{-1}\mathbf{X}^\top\mathbf{Y}}$$

- This corresponds to the solution of the ridge regression (regularized least squares) problem with regularization parameter $\frac{\lambda}{\beta}$

MAP Estimation: An Illustration

$\boldsymbol{w}_{MAP}$ is a compromise between prior's mean and $\boldsymbol{w}_{MLE}$



Summary: MLE vs MAP for Linear Regression

- MLE Objective

$$\arg\max_{\boldsymbol{w}} \log p(\mathbf{Y}|\mathbf{X}, \boldsymbol{w}) = \arg\min_{\boldsymbol{w}} \frac{\beta}{2} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2$$

- MLE solution

$$\boxed{\boldsymbol{w}_{MLE} = (\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top)^{-1} \sum_{n=1}^{N} y_n \boldsymbol{x}_n = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}}$$

- MAP Objective

$$\arg\max_{\boldsymbol{w}} \log p(\boldsymbol{w}|\mathbf{X}, \mathbf{Y}) \propto \arg\max_{\boldsymbol{w}} \log p(\mathbf{Y}|\mathbf{X}, \boldsymbol{w}) p(\boldsymbol{w}) = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \frac{\lambda}{\beta} \boldsymbol{w}^\top \boldsymbol{w}$$

- MAP solution

$$\boxed{\boldsymbol{w}_{MAP} = (\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \frac{\lambda}{\beta} \mathbf{I}_D)^{-1} \sum_{n=1}^{N} y_n \boldsymbol{x}_n = (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta} \mathbf{I}_D)^{-1} \mathbf{X}^\top \mathbf{Y}}$$

The "Fully" Bayesian Approach

- MLE/MAP only provide a point estimate of $\boldsymbol{w}$ (no estimate of uncertainty)
- Let's try to infer the full posterior of $\boldsymbol{w}$: $p(\boldsymbol{w}|\mathbf{X},\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X},\boldsymbol{w})p(\boldsymbol{w})}{p(\mathbf{Y}|\mathbf{X})}$
- Since the likelihood and the prior, both, are Gaussian, the posterior will also be Gaussian (due to conjugacy)
- What will be the posterior's mean and covariance/precision matrix ?
- Since $\mathbf{X}$ is known/fixed, and using the property of Gaussians, given $p(\mathbf{Y}|\mathbf{X},\boldsymbol{w})$ and $p(\boldsymbol{w})$ both Gaussian (refer to the results discussed in lecture 2),

$$
\begin{aligned}
p(\boldsymbol{w}|\mathbf{X},\mathbf{Y}) &= \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \\
\text{where} \quad \boldsymbol{\mu} &= \boldsymbol{\Sigma}(\beta\sum_{n=1}^{N} y_n \boldsymbol{x}_n) = \boldsymbol{\Sigma}(\beta\mathbf{X}^\top\mathbf{Y}) \\
\boldsymbol{\Sigma} &= (\beta\sum_{n=1}^{N} \boldsymbol{x}_n\boldsymbol{x}_n^\top + \lambda\mathbf{I}_D)^{-1} = (\beta\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_D)^{-1}
\end{aligned}
$$

Making Predictions

- MLE and MAP make "plug-in" predictions

$$
\begin{aligned}
p(y_*|\boldsymbol{x}_*, \boldsymbol{w}_{MLE}) &= \mathcal{N}(\boldsymbol{w}_{MLE}^\top \boldsymbol{x}_*, \beta^{-1}) && \text{- MLE prediction} \\
p(y_*|\boldsymbol{x}_*, \boldsymbol{w}_{MAP}) &= \mathcal{N}(\boldsymbol{w}_{MAP}^\top \boldsymbol{x}_*, \beta^{-1}) && \text{- MAP prediction}
\end{aligned}
$$

- MLE/MAP only use a point estimate ($\boldsymbol{w}_{MLE}/\boldsymbol{w}_{MAP}$) for making prediction
- Fully Bayesian approach of making predictions is via the predictive posterior

$$
\boxed{p(y_*|\boldsymbol{x}_*, \mathbf{X}, \mathbf{Y}) = \int_{\boldsymbol{w}} p(y_*|\boldsymbol{x}_*, \boldsymbol{w}) p(\boldsymbol{w}|\mathbf{X}, \mathbf{Y}) d\boldsymbol{w}} \qquad \text{(Predictive Posterior)}
$$

- **Predictive Posterior:** Don't use a single $\boldsymbol{w}$ to make predictions but average $p(y_*|\boldsymbol{x}_*, \boldsymbol{w})$ over all possible $\boldsymbol{w}$'s (each weighted by its posterior probability)
- Since the likelihood $p(y_*|\boldsymbol{x}_*, \boldsymbol{w})$ and posterior $p(\boldsymbol{w}|\mathbf{X}, \mathbf{Y})$ are Gaussian, the predictive posterior is also Gaussian. Thus in the fully Bayesian approach:

$$
\boxed{p(y_*|\boldsymbol{x}_*, \mathbf{X}, \mathbf{Y}) = \mathcal{N}(\boldsymbol{\mu}^\top \boldsymbol{x}_*, \beta^{-1} + \boldsymbol{x}_*^\top \boldsymbol{\Sigma} \boldsymbol{x}_*)}
$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are mean and cov. matrix, resp., of the posterior $p(\boldsymbol{w}|\mathbf{X}, \mathbf{Y})$

Some things we didn't cover..

- How to estimate the model hyperparameters (e.g., precisions β and λ)? The Bayesian approach allows us doing this.
- Nonlinear regression. What to do when a linear model doesn't fit the responses well. Kernel methods (e.g., Gaussian Processes) can handle this.

(We will see these later in the semester)

Next class: Logistic Regression