LaTeX command declarations here.

EECS 445: Machine Learning

Hands On 05: Linear Regression II

- Instructor: **Zhao Fu, Valli, Jacob Abernethy and Jia Deng**
- Date: September 26, 2016

Review: Maximum Likelihood

Suppose we have a set of observed data D and we want to evaluate a parameter setting w:

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

$$p(D) = \sum_w p(D|w)p(w)$$

We call $p(D|w)$ as the likelihood function. Then we have $p(w|D) \propto p(D|w)p(w)$. Suppose $p(w)$ is the same for all w, we can only choose w to maximize likelihood $p(D|w)$, which is to maximize the log-likelihood $\log p(D|w)$.

Review Problem: Maximum Likelihood Estimation

We have observed data $x_1, \cdots, x_n$ drawn from Bernoulli distribution:

$$p(x) = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{if } x = 0 \end{cases}$$

(a) What is the likelihood function based on θ?

(b) What is the log-likelihood function?

(c) Compute estimated θ to maximize the log-likelihood function.

Review: Maximum Likelihood

Suppose we have a set of observed data D and we want to evaluate a parameter setting w:

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

$$p(D) = \sum_w p(D|w)p(w)$$

We call $p(D|w)$ as the likelihood function. Then we have $p(w|D) \propto p(D|w)p(w)$. Suppose $p(w)$ is the same for all w, we can only choose w to maximize likelihood $p(D|w)$, which is to maximize the log-likelihood $\log p(D|w)$.

Solution 1: Maximum Likelihood Estimation

(a)

$$\begin{aligned} L(\theta) &= p(D|\theta) = p(x_1, \ldots, x_n|\theta) \\ &= \textstyle\prod_i p(x_i) = \theta^{\sum 1(x_i=1)}(1-\theta)^{\sum 1(x_i=0)} \\ &= \theta^k(1-\theta)^{n-k} \end{aligned}$$

where k is the number of 1s from the observed data.

(b)

$$\log L(\theta) = k\log(\theta) + (n-k)\log(1-\theta)$$

Solution 1: Maximum Likelihood Estimation

(c) Set the derivative of $log(L(\theta))$ to zero we have

$$\begin{gathered} \frac{\mathrm{d}log(L(\theta))}{\mathrm{d}\theta} = \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0 \\ \frac{k}{\theta} = \frac{n-k}{1-\theta} \\ \theta = \frac{k}{n} \end{gathered}$$

.

Problem 2: Maximum Likelihood Estimation for Gaussian Distribution

We have observed data $x_1, \cdots, x_n$ drawn from Normal distribution:

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

(a) What is the likelihood function based on μ and σ^2?

(b) What is the log-likelihood function?

(c) Compute estimated parameters μ and σ^2 to maximize the log-likelihood function.

Solution 2

We have observed data $x_1, \cdots, x_n$ drawn from Normal distribution:

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

(a) What is the log-likelihood function?

Answer: $-(n/2)\log\sigma - \sum_{i=1}^{n} \frac{1}{2\sigma^2}(x_i - \mu)^2$

(b) Compute estimated parameters μ and σ^2 to maximize the log-likelihood function.

Answer:

- $\mu_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- $\sigma^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{\text{MLE}})^2$

Regularized Linear Regression

Regularized Least Squares: Objective Function

- Recall the objective function we minimizes in last lecture is

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{w}^T\phi(\mathbf{x}_n) - t_n\right)^2$$

- To penalize the large coefficients, we will add one penalization/regularization term to it and minimize them altogether.

$$E(\mathbf{w}) = \underbrace{\frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{w}^T\phi(\mathbf{x}_n) - t_n\right)^2}_{E_D(\mathbf{w})} + \underbrace{\boxed{\frac{\lambda}{2}\|\mathbf{w}\|^2}}_{E_W(\mathbf{w})}$$

  of which $E_D(\mathbf{w})$ represents the term of sum of squared errors and $E_W(\mathbf{w})$ is the regularization term.
- λ is the regularization coefficient.
- If λ is large, $E_{\mathbf{W}}(\mathbf{w})$ will dominate the objective function. As a result we will focus more on minimizing $E_W(\mathbf{w})$ and the resulting solution $\mathbf{w}$ tends to have smaller norm and the $E_D(\mathbf{w})$ term will be larger.

Regularized Least Squares: Derivation

- Based on what we have derived in last lecture, we could write the objective function as

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{w}^T\phi(\mathbf{x}_n) - t_n\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

Exercise: Derive the gradient in element-wise to verify the above result, i.e. using $\phi(\mathbf{x}_n)_d$ and w_d to represent $E(w_1, w_2, \ldots, w_D)$ and derive $\frac{\partial E}{\partial w_d}$. Suppose $\phi(\mathbf{x_n}) \in \mathbb{R}^D$.

Regularized Least Squares: Solution

- Based on what we have derived in last lecture, we could write the objective function as

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(\sum_{d=1}^{D} w_d\phi_d(\mathbf{x}_n) - t_n)^2 + \frac{\lambda}{2}\sum_{d=1}^{D} w_d^2$$

$$\frac{\partial E}{\partial w_d} = \sum_{n=1}^{N}\phi_d(\mathbf{x}_n)(\sum_{d=1}^{D} w_d\phi_d(\mathbf{x}_n) - t_n) + \lambda w_d$$

$$\frac{\partial E}{\partial w_d} = \sum_{n=1}^{N}\phi_d(\mathbf{x}_n)(\mathbf{w}^T\phi(\mathbf{x}_n) - t_n) + \lambda w_d$$

- The gradient is

$$\begin{aligned}\nabla_{\mathbf{w}} E(\mathbf{w}) &= \mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w} - \mathbf{\Phi}^T\mathbf{t} + \lambda\mathbf{w} \\ &= (\mathbf{\Phi}^T\mathbf{\Phi} + \lambda I)\mathbf{w} - \mathbf{\Phi}^T\mathbf{t}\end{aligned}$$

- Setting the gradient to 0, we will get the solution

$$\boxed{\hat{\mathbf{w}} = (\mathbf{\Phi}^T\mathbf{\Phi} + \lambda I)^{-1}\mathbf{\Phi}^T\mathbf{t}}$$

Regularized Least Squares: Closed Form

In the solution to ordinary least squares which is $\hat{\mathbf{w}} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{t}$, we cannot guarantee $\mathbf{\Phi}^T\mathbf{\Phi}$ is invertible. But in regularized least squares, if $\lambda > 0$, $\mathbf{\Phi}^T\mathbf{\Phi} + \lambda I$ is always invertible.

Exercise: To be invertible, a matrix needs to be full rank. Argue that $\mathbf{\Phi}^T\mathbf{\Phi} + \lambda I$ is full rank by characterizing its p eigenvalues in terms of the singular values of $\mathbf{\Phi}$ and λ.

Solution:

Suppose $\mathbf{\Phi} = U^T\Lambda V$ which is SVD of $\mathbf{\Phi}$, we have $\mathbf{\Phi}^T\mathbf{\Phi} = V^T\Lambda^2 V$.

Then we have $(\mathbf{\Phi}^T\mathbf{\Phi} + \lambda I)V^T = V^T(\Lambda^2 + \lambda I)$.

The i^{th} eigenvalue of $(\mathbf{\Phi}^T\mathbf{\Phi} + \lambda I)$ is $\lambda_i^2 + \lambda > 0$ where λ_i is the singular value for $\mathbf{\Phi}$.

Then $\det(\mathbf{\Phi}^T\mathbf{\Phi} + \lambda I) = \prod(\lambda_i^2 + \lambda) > 0$, which means $\mathbf{\Phi}^T\mathbf{\Phi} + \lambda I$ is invertable.

Regularized Least Squares: Different Norms

- The ℓ^p norm of a vector $\mathbf{x}$ is defined as

$$\|\mathbf{x}\|_p = (\sum_{j=1}^{M} |x_j|^p)^{\frac{1}{p}}$$

- For the regularized least squares above, we used ℓ^2 norm. We could also use other ℓ^p norms for different regularizers and the objective function becomes

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{w}^T\phi(\mathbf{x}_n) - t_n\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|_p^p$$

Exercise: Derive the element-wise gradient for the above ℓ^p norm regularized energy function.

Regularized Least Squares: Summary

- Simple modification of linear regression
- ℓ^2 Regularization controls the tradeoff between *fitting error* and *complexity*.
  - Small ℓ^2 regularization results in complex models, but with risk of overfitting
  - Large ℓ^2 regularization results in simple models, but with risk of underfitting
- It is important to find an optimal regularization that *balances* between the two

Probablistic Interpretation of Least Squares Regression

- We have showed derived the solution to least squares regression by minimizing objective function. Now we will provide a probablistic perspective. Specifically, we will show the solution to **regular least squares** is just the **maximum likelihood** estimate of $\mathbf{w}$ and the solution to **regularized least squares** is the **Maximum a Posteriori** estimate.

Some Background

- Gaussian Distribution

$$\mathcal{N}(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{(x-\mu)^2}{2\sigma^2}\right]$$

- **Maximum Likelihood Estimation** and **Maximum a Posteriori Estimation (MAP)**
  - For distribution $t \sim p(t|\theta)$. θ is some unknown parameter (like mean or variance) to be estimated.
  - Given observation $\mathbf{t} = (t_1, t_2, \ldots, t_N)$,
    - The Maximum Likelihood Estimator is

$$\theta_{ML} = \arg\max \prod_{n=1}^{N} p(t_n|\theta)$$

    - If we have some prior knowledge about θ, the MAP estimator is

$$\theta_{MAP} = \arg\max \prod_{n=1}^{N} p(\theta|t_n) \quad (\text{Posteriori Probability of } \theta)$$

Maximum Likelihood Estimator $\mathbf{w}_{ML}$

- We assume the **signal+noise** model of single data $(\mathbf{x}, t)$ is

$$t = \mathbf{w}^T \phi(\mathbf{x}) + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \beta^{-1})$$

  of which $\mathbf{w}^T \phi(\mathbf{x})$ is the true model, ϵ is the perturbation/randomness.
- Since $\mathbf{w}^T \phi(\mathbf{x})$ is deterministic/non-random, we have

$$t \sim \mathcal{N}(\mathbf{w}^T \phi(\mathbf{x}), \beta^{-1})$$

Exercise:

- Derive the likelihood function for a single data $p(t_n|\mathbf{x}_n, \mathbf{w}, \beta)$.
- Derive the complete log likelihood function for the whole dataset $\ln p(\mathbf{t}|\mathcal{X}, \mathbf{w}, \beta)$.
- Using maximum likelihood to estimate parameter $\mathbf{w}$.

Maximum Likelihood Estimator $\mathbf{w}_{ML}$

- The **likelihood function** of t is just **probability density function (PDF)** of t

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|\mathbf{w}^T\phi(\mathbf{x}), \beta^{-1})$$

- For inputs $\mathcal{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ and target values $\mathbf{t} = (t_1, \dots, t_n)$, the data likelihood is

$$p(\mathbf{t}|\mathcal{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\phi(\mathbf{x}_n), \beta^{-1})$$

- **Notation Clarification**
  - $p(t|x, w, \beta)$ it the PDF of t whose distribution is parameterized by $x, \mathbf{w}, \beta$.
  - $\mathcal{N}(\mathbf{w}^T\phi(\mathbf{x}), \beta^{-1})$ is Gaussian distribution with **mean** $\mathbf{w}^T\phi(\mathbf{x})$ and **variance** β^{-1}.
  - $\mathcal{N}(t|\mathbf{w}^T\phi(\mathbf{x}), \beta^{-1})$ is the PDF of $\mathbf{t}$ which has Gaussian distribution $\mathcal{N}(\mathbf{w}^T\phi(\mathbf{x}), \beta^{-1})$

Maximum Likelihood Estimator $\mathbf{w}_{ML}$: Derivation

- Single data likelihood is

$$p(t_n|\mathbf{x}_n, \mathbf{w}, \beta) = \mathcal{N}(t_n|\mathbf{w}^T\phi(\mathbf{x}_n), \beta^{-1}) = \frac{1}{\sqrt{2\pi\beta^{-1}}}\exp\left\{-\frac{1}{2\beta^{-1}}(t_n - \mathbf{w}^T\phi(x_n))^2\right\}$$

- Single data log-likelihood is

$$\ln p(t_n|\mathbf{x}_n, \mathbf{w}, \beta) = -\frac{1}{2}\ln 2\pi\beta^{-1} - \frac{\beta}{2}(\mathbf{w}^T\phi(x_n) - t_n)^2$$

  We use logarithm because maximizer of $f(x)$ is the same as maximizer of $\log f(x)$. Logarithm can convert product to summation which makes life easier.
- Complete data log-likelohood is

$$\begin{aligned}\ln p(\mathbf{t}|\mathcal{X}, \mathbf{w}, \beta) &= \ln\left[\prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \beta)\right] = \sum_{n=1}^{N} \ln p(t_n|\mathbf{x}_n, \mathbf{w}, \beta) \\ &= \sum_{n=1}^{N}\left[-\frac{1}{2}\ln 2\pi\beta^{-1} - \frac{\beta}{2}(\mathbf{w}^T\phi(x_n) - t_n)^2\right]\end{aligned}$$

- Maximum likelihood estimate $\mathbf{w}_{ML}$ is

$$\begin{aligned}\mathbf{w}_{ML} &= \arg\max_{\mathbf{w}} \ln p(\mathbf{t}|\mathcal{X}, \mathbf{w}, \beta) \\ &= \arg\max_{\mathbf{w}} \sum_{n=1}^{N} \left[-\frac{1}{2}\ln 2\pi\beta^{-1} - \frac{\beta}{2}(\mathbf{w}^T\phi(x_n) - t_n)^2 \right] \\ &= \arg\max_{\mathbf{w}} \sum_{n=1}^{N} \left[-\frac{\beta}{2}(\mathbf{w}^T\phi(x_n) - t_n)^2 \right] \\ &= \arg\min_{\mathbf{w}} \sum_{n=1}^{N} \left[(\mathbf{w}^T\phi(x_n) - t_n)^2 \right]\end{aligned}$$

- Familiar? Recall the objective function we minimized in least squares is $E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{w}^T\phi(\mathbf{x}_n) - t_n\right)^2$, so we could conclude that

$$\boxed{\mathbf{w}_{ML} = \hat{\mathbf{w}}_{LS} = \mathbf{\Phi}^{\dagger}\mathbf{t}}$$

MAP Estimator $\mathbf{w}_{MAP}$

- The **MAP estimator** is obtained by

$$\begin{aligned}\mathbf{w}_{MAP} &= \arg\max p(\mathbf{w}|\mathbf{t}, \mathcal{X}, \beta) && (\text{Posteriori Probability}) \\ &= \arg\max \frac{p(\mathbf{w}, \mathbf{t}, \mathcal{X}, \beta)}{p(\mathcal{X}, t, \beta)} \\ &= \arg\max \frac{p(\mathbf{t}|\mathbf{w}, \mathcal{X}, \beta)p(\mathbf{w}, \mathcal{X}, \beta)}{p(\mathcal{X}, t, \beta)} \\ &= \arg\max p(\mathbf{t}|\mathbf{w}, \mathcal{X}, \beta)p(\mathbf{w}, \mathcal{X}, \beta) && (p(X, t, \beta) \text{ is irrelevant to } \mathbf{w}) \\ &= \arg\max p(\mathbf{t}|\mathbf{w}, \mathcal{X}, \beta)p(\mathbf{w})p(\mathcal{X})p(\beta) && (\text{Independence}) \\ &= \arg\max p(\mathbf{t}|\mathbf{w}, \mathcal{X}, \beta)p(\mathbf{w}) && (\text{Likelihood} \times \text{Prior})\end{aligned}$$

  We are just using **Bayes Theorem** for the above steps.
- The only difference from ML estimator is we have an extra term of PDF of $\mathbf{w}$. This is the **prior belief** of $\mathbf{w}$. Here, we assume,

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}I)$$

Exercise: Derive the MAP Estimator of $\mathbf{w}$ and compare the solution with regularized linear regression. What is $\boldsymbol{\lambda}$ in this case?

MAP Estimator $\mathbf{w}_{MAP}$: Derivation

- $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha^{-1} I)$ is **multivariate Gaussian** which has PDF

$$p(\mathbf{w}) = \frac{1}{\left(\sqrt{2\pi\alpha^{-1}}\right)^N} \exp\left\{-\frac{1}{2\alpha^{-1}} \sum_{n=1}^{N} w_n^2\right\}$$

- So the MAP estimator is

$$\begin{aligned}
\mathbf{w}_{MAP} &= \arg\max_{\mathbf{w}} p(\mathbf{t}|\mathbf{w}, \mathcal{X}, \beta) p(\mathbf{w}) = \arg\max_{\mathbf{w}} \left[\ln p(\mathbf{t}|\mathbf{w}, \mathcal{X}, \beta) + \ln p(\mathbf{w})\right] \\
&= \arg\min_{\mathbf{w}} \left[\sum_{n=1}^{N} \frac{\beta}{2} (\mathbf{w}^T \phi(x_n) - t_n)^2 + \frac{\alpha}{2} \sum_{n=1}^{N} w_n^2\right] \\
&= \arg\min_{\mathbf{w}} \left[\sum_{n=1}^{N} \frac{1}{2} (\mathbf{w}^T \phi(x_n) - t_n)^2 + \frac{1}{2} \frac{\alpha}{\beta} \|\mathbf{w}\|^2\right]
\end{aligned}$$

- Exactly the objective in regularized least squares! So

$$\boxed{\mathbf{w}_{MAP} = \hat{\mathbf{w}} = \left(\mathbf{\Phi}^T \mathbf{\Phi} + \frac{\alpha}{\beta} I\right)^{-1} \mathbf{\Phi}^T \mathbf{t}}$$

- Compared with ℓ^2 norm regularized least square, we have $\boldsymbol{\lambda} = \frac{\alpha}{\beta}$.

Problem 5a: MAP estimation for Linear Regression with unusual Prior

Assume we have n vectors $\mathbf{x}_1, \cdots, \mathbf{x}_n$. We also assume that for each $\mathbf{x}_i$ we have observed a *target* value t_i, where

$$\begin{aligned}
t_i &= \mathbf{w}^T \mathbf{x_i} + \epsilon \\
\epsilon &\sim \mathcal{N}(0, \beta^{-1})
\end{aligned}$$

where ϵ is the "noise term".

(a) Quick quiz: what is the likelihood *given* $\mathbf{w}$? That is, what's $p(t_i|\mathbf{x}_i, \mathbf{w})$?

Answer: $p(t_i|\mathbf{x}_i, \mathbf{w}) = \mathcal{N}(t_i|\mathbf{w}^\top \mathbf{x_i}, \beta^{-1}) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{\beta}{2}(t_i - \mathbf{w}^\top \mathbf{x_i})^2\right)$

Problem 5: MAP estimation for Linear Regression with unusual Prior

Assume we have n vectors $\mathbf{x}_1, \cdots, \mathbf{x}_n$. We also assume that for each $\mathbf{x}_i$ we have observed a *target* value t_i, sampled IID. We will also put a *prior* on $\mathbf{w}$, using PSD matrix Σ.

$$t_i = \mathbf{w}^T \mathbf{x_i} + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \beta^{-1})$$
$$\mathbf{w} \sim \mathcal{N}(0, \Sigma)$$

Note: the difference here is that our prior is a multivariate gaussian with *non-identity* covariance! Also we let $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$

(a) Compute the log posterior function, $\log p(\mathbf{w}|\mathbf{t}, \mathcal{X}, \beta)$

Hint: use Bayes' Rule

(b) Compute the MAP estimate of $\mathbf{w}$ for this model

Hint: the solution is very similar to the MAP estimate for a gaussian prior with identity covariance