Machine Learning

Lecture 4: Regularization and Bayesian Statistics

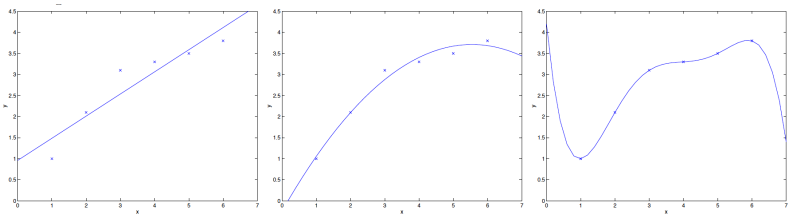
Feng Li

fli@sdu.edu.cn
https://funglee.github.io

School of Computer Science and Technology
Shandong University

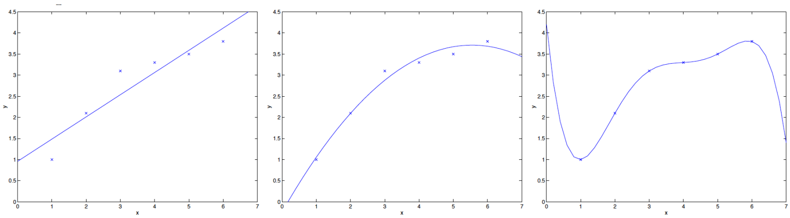
Fall 2018

Overfitting Problem



- $y = \theta_0 + \theta_1 x$
- $y = \theta_0 + \theta_1 x + \theta_2 x^2$
- $y = \theta_0 + \theta_1 x + \cdots + \theta_5 x^5$

Overfitting Problem (Contd.)



- Underfitting, or high bias, is when the form of our hypothesis function h maps poorly to the trend of the data
- Overfitting, or high variance, is caused by a hypothesis function that fits the available data but does not generalize well to predict new data

Addressing The Overfitting Problem

- Reduce the number of features
    - Manually select which features to keep
    - Use a model selection algorithm
- Regularization
    - Keep all the features, but reduce the magnitude of parameters θ_j
    - Regularization works well when we have a lot of slightly useful features.

Optimizing Cost Function by Regularization

- Consider the following function

$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

- To eliminate the influence of $\theta_3 x^3$ and $\theta_4 x^4$ to smoothen hypothesis function, the cost function can be modified as follows

$$\min_{\theta} \frac{1}{2m} \left[\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + 1000 \cdot \theta_3^2 + 1000 \cdot \theta_4^2 \right]$$

- Large values of θ_3 and θ_4 will increase the objective function

Optimizing Cost Function by Regularization (Contd.)

- A more general form

$$\min_{\theta} \frac{1}{2m} \left[\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2 \right]$$

  where λ is the regularization parameter
- As the magnitudes of the fitting parameters increase, there will be an increasing penalty on the cost function
- This penalty is dependent on the squares of the parameters as well as the magnitude of λ

Regularized Linear Regression

- Gradient descent
  - **Repeat** $\{$

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_0^{(i)}$$

$$\theta_j := \theta_j - \alpha \left[(\frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}) + \frac{\lambda}{m} \theta_j \right]$$

$\}$ **until** convergence condition is satisfied

- The regularization is performed by $\lambda\theta_j/m$

Regularized Linear Regression (Contd.)

- Normal equation

$$\theta = (X^T X + \lambda \cdot L)^{-1} X^T \vec{y}$$

where

$$L = \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$$

Regularized Logistic Regression

- Recall the cost function for logistic regression

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))]$$

- Adding a term for regularization

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

- Gradient descent:
  Repeat
  - $\theta_0 := \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_0^{(i)}$
  - $\theta_j := \theta_j - \alpha\left(\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)} + \frac{\lambda}{m}\theta_j\right)$ for $j = 1, 2, \cdots, n$

  until convergence condition is satisfied

Parameter Estimation in Probabilistic Models

- Assume data are generated via probabilistic model

$$d \sim p(d; \theta)$$

- $p(d; \theta)$: Probability distribution underlying the data
  - θ: Fixed but unknown distribution parameter
- Given: m independent and identically distributed (i.i.d.) samples of the data

$$D = \{d^{(i)}\}_{i=1,\cdots,m}$$

- Independent and Identically Distributed
  - Given θ, each sample is independent of all other samples
  - All samples drawn from the same distribution
- Goal: Estimate parameter θ that best models/describes the data
- Several ways to define the "best"

Maximum Likelihood Estimation (MLE)

- Maximum Likelihood Estimation (MLE): Choose the parameter θ that maximizes the probability of the data, given that parameter
- Probability of the data, given the parameters, is called *likelihood*, a function of θ and defined as:

$$L(\theta) = p(D; \theta) = \prod_{i=1}^{m} p(d^{(i)}; \theta)$$

- MLE typically maximizes the *log-likelihood* instead of the likelihood

$$\ell(\theta) = \log L(\theta) = \log \prod_{i=1}^{m} p(d^{(i)}; \theta) = \sum_{i=1}^{m} \log p(d^{(i)}; \theta)$$

- Maximum likelihood parameter estimation

$$\theta_{MLE} = \arg\max_{\theta} \ell(\theta) = \arg\max_{\theta} \sum_{i=1}^{m} \log p(d^{(i)}; \theta)$$

Maximum-a-Posteriori Estimation (MAP)

- Maximum-a-Posteriori Estimation (MAP): Maximize the posterior probability of θ (i.e., probability in the light of the observed data)
- Posterior probability of θ is given by the Bayes Rule

$$p(\theta \mid D) = \frac{p(\theta)p(D \mid \theta)}{p(D)}$$

- $p(\theta)$: Prior probability of θ (without having seen any data)
- $p(D)$: Probability of the data (independent of θ)

$$p(D) = \int_{\theta} p(\theta)p(D \mid \theta)d\theta$$

- The Bayes Rule lets us update our belief about θ in the light of observed data
- While doing MAP, we usually maximize the log of the posteriori probability

Maximum-a-Posteriori Estimation (Contd.)

- Maximum-a-Posteriori parameter estimation

$$
\begin{aligned}
\theta_{MAP} &= \arg\max_{\theta} p(\theta \mid D) \\
&= \arg\max_{\theta} \frac{p(\theta)p(D \mid \theta)}{p(D)} \\
&= \arg\max_{\theta} p(\theta)p(D \mid \theta) \\
&= \arg\max_{\theta} \log p(\theta)p(D \mid \theta) \\
&= \arg\max_{\theta} (\log p(\theta) + \log p(D \mid \theta)) \\
&= \arg\max_{\theta} \left(\log p(\theta) + \sum_{i=1}^{m} \log p(d^{(i)} \mid \theta) \right)
\end{aligned}
$$

Maximum-a-Posteriori Estimation (Contd.)

- Same as MLE except the extra log-prior-distribution term!
- MAP allows incorporating our prior knowledge about θ in its estimation

$$\begin{aligned}\theta_{MLE} &= \arg\max_{\theta} \ell(\theta) = \arg\max_{\theta} \sum_{i=1}^{m} \log p(d^{(i)}; \theta) \\ \theta_{MAP} &= \arg\max_{\theta} \left(\log p(\theta) + \sum_{i=1}^{m} \log p(d^{(i)} \mid \theta) \right)\end{aligned}$$

Linear Regression: MLE Solution

- For each $(x^{(i)}, y^{(i)})$,
$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$
- The noise $\epsilon^{(i)}$ is drawn from a Gaussian distribution
$$\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$$
- Each $y^{(i)}$ is drawn from the following Gaussian
$$y^{(i)} | x^{(i)}; \theta \sim \mathcal{N}(\theta^T x^{(i)}, \sigma^2)$$
- The log-likelihood
$$\ell(\theta) = \log L(\theta) = m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{\sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}$$
- Maximize $\ell(\theta)$
$$\theta_{MLE} = \arg\min_{\theta} \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2$$

Linear Regression: MAP Solution

- θ follows a Gaussian distribution $\theta \sim \mathcal{N}(0, \sigma^2 I)$

$$p(\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\theta^T\theta}{2\sigma^2}\right)$$

and thus

$$\log p(\theta) = n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{\theta^T\theta}{2\sigma^2}$$

- The MAP solution

$$\theta_{MAP} = \arg\max_\theta \{\sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)}, \theta) + \log p(\theta)\}$$

$$= \arg\max_\theta \left\{ m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{\sum_{i=1}^m (y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2} + n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{\theta^T\theta}{2\sigma^2} \right\}$$

$$= \arg\min_\theta \left\{ \frac{\sum_{i=1}^m (y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2} + \frac{\theta^T\theta}{2\sigma^2} \right\}$$

Linear Regression: MLE vs MAP

- MLE solution

$$\theta_{MLE} = \arg\min_{\theta} \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2$$

- MAP solution

$$\theta_{MAP} = \arg\min_{\theta} \left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2 + \frac{\theta^T \theta}{2\sigma^2} \right\}$$

- What do we learn?
  - MLE estimation of a parameter leads to unregularized solution
  - MAP estimation of a parameter leads to regularized solution
  - The prior distribution acts as a regularizer in MAP estimation
- For MAP, different prior distributions lead to different regularizers
  - Gaussian prior on θ regularizes the ℓ_2 norm of θ
  - Laplace prior $\exp(C\|\theta\|_1)$ on θ regularizes the the ℓ_1 norm of θ

Probabilistic Classification: Logistic Regression

- Often we do not just care about predicting the label y for an example
- Rather, we want to predict the label probabilities $p(y|x, \theta)$
- Consider the following function ($y \in \{-1, 1\}$)

$$p(y \mid x; \theta) = g(y\theta^T x) = \frac{1}{1 + \exp(-y\theta^T x)}$$

- g is the logistic function which maps all real number into $(0, 1)$
- This is logistic regression model, which is a classification model

Logistic Regression

- What does the decision boundary look like for Logistic Regression?
- At the decision boundary labels +1/-1 becomes equiprobable

$$
\begin{aligned}
p(y=+1 \mid x, \theta) &= p(y=-1 \mid x, \theta) \\
\frac{1}{1+\exp(-\theta^T x)} &= \frac{1}{1+\exp(\theta^T x)} \\
\exp(-\theta^T x) &= \exp(\theta^T x) \\
\theta^T x &= 0
\end{aligned}
$$

- The decision boundary is therefore linear $\Rightarrow$ Logistic Regression is a linear classifier (note: it is possible to kernelize and make it nonlinear)

Logistic Regression: MLE Solution

- Log-likelihood

$$\begin{aligned}
\ell(\theta) &= \log \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}, \theta) \\
&= \sum_{i=1}^{m} \log p(y^{(i)} \mid x^{(i)}, \theta) \\
&= \sum_{i=1}^{m} \log \frac{1}{1 + \exp(-y^{(i)} \theta^T x^{(i)})} \\
&= - \sum_{i=1}^{m} \log[1 + \exp(-y^{(i)} \theta^T x^{(i)})]
\end{aligned}$$

- MLE solution

$$\theta_{MLE} = \arg\min_{\theta} \sum_{i=1}^{m} \log[1 + \exp(-y^{(i)} \theta^T x^{(i)})]$$

- No close-form solution exists, but we can do gradient descent on θ

Logistic Regression: MAP Solution

- Again, assume θ follows a Gaussian distribution $\theta \sim \mathcal{N}(0, \sigma^2 I)$

$$p(\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\theta^T\theta}{2\sigma^2}\right) \quad \Rightarrow \quad \log p(\theta) = -n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{\theta^T\theta}{2\sigma^2}$$

- MAP solution

$$\theta_{MAP} = \arg\min_{\theta} \sum_{i=1}^{m} \log[1 + \exp(-y^{(i)}\theta^T x^{(i)})] + \frac{1}{2\sigma^2}\theta^T\theta$$

- See "A comparison of numerical optimizers for logistic regression" by Tom Minka on optimization techniques (gradient descent and others) for logistic regression (both MLE and MAP)

Logistic Regression: MLE vs MAP

- MLE solution

$$\theta_{MLE} = \arg\min_{\theta} \sum_{i=1}^{m} \log[1 + \exp(-y^{(i)}\theta^T x^{(i)})]$$

- MAP solution

$$\theta_{MAP} = \arg\min_{\theta} \sum_{i=1}^{m} \log[1 + \exp(-y^{(i)}\theta^T x^{(i)})] + \frac{1}{2\sigma^2}\theta^T\theta$$

- Take-home messages (we already saw these before :-))
  - MLE estimation of a parameter leads to unregularized solutions
  - MAP estimation of a parameter leads to regularized solutions
  - The prior distribution acts as a regularizer in MAP estimation
- For MAP, different prior distributions lead to different regularizers
  - Gaussian prior on θ regularizes the ℓ_2 norm of θ
  - Laplace prior $\exp(C\|\theta\|_1)$ on θ regularizes the the ℓ_1 norm of θ

Thanks!

Q & A