12110817 张展玮

Solution 1

a)

$$X = (A^T Q^{-1} A)^{-1} A^T Q^{-1} Y$$

b)

$$L(X, \lambda) = \frac{1}{2}(Y - AX)^T Q^{-1}(Y - AX) + \lambda(b^T X - c)$$

$$\frac{\partial}{\partial X} L(X, \lambda) = -(Y - AX)^T Q^{-1} A + \lambda b^T = 0$$

$$\frac{\partial}{\partial \lambda} L(X, \lambda) = b^T X - c = 0$$

$$X = (A^T Q^{-1} A)^{-1}(A^T Q^{-1} Y - \frac{1}{2}\lambda b)$$

where λ is the Lagrange multiplier. Then by solving the two equations to get X.

c)

$$\mathcal{L}(\mathbf{X}, \lambda_1, \lambda_2) = (\mathbf{Y} - \mathbf{AX})^\top \mathbf{Q}^{-1}(\mathbf{Y} - \mathbf{AX}) + \lambda_1(\mathbf{b}^\top \mathbf{X} - c) + \lambda_2(\mathbf{X}^\top \mathbf{X} - d)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{X}} = -2\mathbf{A}^\top \mathbf{Q}^{-1}(\mathbf{Y} - \mathbf{AX}) + \lambda_1 \mathbf{b} + 2\lambda_2 \mathbf{X} = 0$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_1} = \mathbf{b}^\top \mathbf{X} - c = 0$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_2} = \mathbf{X}^\top \mathbf{X} - d = 0$$

Then by solving the two equations to get X.

d)

If both A and X are unknown, solving them requires an iterative method due to the joint estimation problem. This can be approached by using alternating minimization.

First, to solve for X with a fixed A,

$$\mathcal{L}(X, \lambda) = (Y - AX)^\top Q^{-1}(Y - AX) + \lambda(X^\top X - d)$$

$$\frac{\partial \mathcal{L}}{\partial X} = -2A^\top Q^{-1}(Y - AX) + 2\lambda X = 0$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = X^\top X - d = 0$$

Solving these equations will yield the value for X.

Next, fix X and solve for A,

$$\mathcal{L}(A, \lambda') = (Y - AX)^\top Q^{-1}(Y - AX) + \lambda'(\text{Trace}(A^\top A) - e)$$
$$\frac{\partial \mathcal{L}}{\partial A} = -2Q^{-1}(Y - AX)X^\top + 2\lambda' A = 0$$
$$\frac{\partial \mathcal{L}}{\partial \lambda'} = \text{Trace}(A^\top A) - e = 0$$

Solving these equations will yield the value for A.

By iterating between these two steps, adjusting X with a fixed A and then adjusting A with the new X, and repeating until the change in the values of X and A falls below a certain threshold, we find the optimal solution to the least squares problem with the given constraints.

Solution 2

Conditional distribution

$$p(Y|X) = \mathcal{N}(\boldsymbol{Y}|AX, \beta^{-1}\boldsymbol{I})$$

Joint distribution

$$p(Y, X) = \mathcal{N}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \middle| \begin{bmatrix} m_0 \\ Am_0 \end{bmatrix}, \begin{bmatrix} \Sigma_0 & \Sigma_0 A^\top \\ A\Sigma_0 & A\Sigma_0 A^\top + \beta^{-1} I \end{bmatrix}\right)$$

Marginal distribution

$$p(Y) = \int p(Y, X) dX = \mathcal{N}(Y|Am_0, A\Sigma_0 A^\top + \beta^{-1} I)$$

Posterior distribution

$$\begin{aligned} p(X|Y = \boldsymbol{y}, \beta, \boldsymbol{m}_0, \boldsymbol{\Sigma}_0) &\propto p(Y = y|X, \beta)p(X|\boldsymbol{m}_0, \boldsymbol{\Sigma}_0) \\ &= \mathcal{N}(X|\mu_{X|Y}, \Sigma_{X|Y}) \end{aligned}$$

$$\mu_{X|Y} = \mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(y - \mu_Y) = m_0 + \Sigma_0 A^T (A\Sigma_0 A^T + \beta^{-1} I)^{-1}(y - Am_0)$$
$$\Sigma_{X|Y} = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} = \Sigma_0 - \Sigma_0 A^T (A\Sigma_0 A^T + \beta^{-1} I)^{-1} A\Sigma_0$$

Posterior predictive distribution

$$\begin{aligned} p(\tilde{Y}|Y = y, \beta, m_0, \Sigma_0) &= \int p(\tilde{Y}|X)p(X|Y = y, \beta, m_0, \Sigma_0)dX \\ &= \mathcal{N}(X|\mu_{X|Y}, \Sigma_{X|Y}) \times \mathcal{N}(\boldsymbol{Y}|AX, \beta^{-1}\boldsymbol{I}) \end{aligned}$$

Prior predictive distribution

$$p(Y|\beta, m_0, \Sigma_0) = \mathcal{N}(Y|Am_0, A\Sigma_0 A^\top + \beta^{-1} I)$$

Solution 3

The posterior distribution is

$$p(\boldsymbol{w}|\mathcal{D},\beta,\boldsymbol{m}_0,\alpha)=\mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N,\boldsymbol{S}_N)$$

where

$$\begin{aligned}\boldsymbol{m}_N&=\boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0+\beta\boldsymbol{\Phi}^T\boldsymbol{t})\\ \boldsymbol{S}_N^{-1}&=\boldsymbol{S}_0^{-1}+\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\\ \boldsymbol{S}_0&=\alpha^{-1}\boldsymbol{I}\end{aligned}$$

The posterior predictive distribution is

$$p(\hat{y}|\hat{x},\mathcal{D},\beta,\boldsymbol{m}_0,\alpha)=\mathcal{N}(\hat{y}|\boldsymbol{m}_N^T\boldsymbol{\phi}(\hat{x}),\sigma_N^2(\hat{x}))$$

where

$$\sigma_N^2(\hat{x})=\frac{1}{\beta}+\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{S}_N\boldsymbol{\phi}(\boldsymbol{x})$$

The prior predictive distribution is

$$p(\mathcal{D}|\beta,\boldsymbol{m}_0,\alpha)=\prod_{n=1}^{n}\mathcal{N}(y_n|\phi_n^Tm_0,\phi_n^T\alpha^{-1}\phi_n+\beta^{-1})$$

Solution 4. Logistics Regression

The posterior distribution

$$p(\boldsymbol{w}|\mathcal{D},\boldsymbol{m}_0,\alpha)\propto p(\boldsymbol{w}|\boldsymbol{m}_0,\alpha)p(\mathcal{D}|\boldsymbol{w})=p(\boldsymbol{w}|\boldsymbol{m}_0,\alpha)\prod_{n=1}^{N}p(t_n|\boldsymbol{w})$$

$$\ln p(\boldsymbol{w}|\mathcal{D},\boldsymbol{m}_0,\alpha)=-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_0)^T\boldsymbol{S}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)+\sum_{i=1}^{N}\{t_n\ln y_n+(1-t_n)\ln(1-y_n)\}+C$$

$$S_N^{-1}=-\nabla\nabla\ln p(\boldsymbol{w}|\mathcal{D},\boldsymbol{m}_0,\alpha)=\boldsymbol{S}_0^{-1}+\sum_{n=1}^{N}y_n(1-y_n)\phi_n\phi_n^T$$

Hence,

$$q(\boldsymbol{w})=\mathcal{N}(\boldsymbol{w}|\boldsymbol{w}_{MAP},\boldsymbol{S}_N)$$

The posterior predictive distribution

$$p(t|x,\mathcal{D},\boldsymbol{m}_0,\alpha)$$

The prior predictive distribution

$$p(D|\boldsymbol{m}_0,\alpha)=\prod_{n=1}^{N}p(t_n|\boldsymbol{x}_n,\boldsymbol{m}_0,\alpha)$$

Solution 5

(1)

$$\frac{\partial y}{\partial w^{(1)}}=\frac{\partial y}{\partial a_2}\cdot\frac{\partial a_2}{\partial z}\cdot\frac{\partial z}{\partial a_1}\cdot\frac{\partial a_1}{\partial w^{(1)}}=y(1-y)w^{(2)}h'(a_1)x$$

$$\frac{\partial y}{\partial w^{(2)}} = \frac{\partial y}{\partial a_2} \cdot \frac{\partial a_2}{\partial w^{(2)}} = y(1-y)z$$

$$\frac{\partial y}{\partial a_1} = \frac{\partial y}{\partial a_2} \cdot \frac{\partial a_2}{\partial z} \cdot \frac{\partial z}{\partial a_1} = y(1-y)w^{(2)}h'(a_1)$$

$$\frac{\partial y}{\partial a_2} = \frac{\partial y}{\partial a_2} = y(1-y)$$

$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial a_2} \cdot \frac{\partial a_2}{\partial z} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial a_1}{\partial x} = y(1-y)w^{(2)}h'(a_1)w^{(1)}$$

(2)

Use MSE

$$L = \frac{1}{2}(y-t)^2$$

$$\begin{aligned}
\Delta w^{(2)} &= -\eta \frac{\partial L}{\partial w^{(2)}} \\
&= -\eta \frac{\partial L}{\partial y} \frac{\partial y}{\partial a_2} \frac{\partial a_2}{\partial w^{(2)}} \\
&= \eta(t-y)y'(a_2)z \\
&= \eta(t-y)y(1-y)z
\end{aligned}$$

$$\begin{aligned}
\Delta w^{(1)} &= -\eta \frac{\partial L}{\partial w^{(1)}} \\
&= -\eta \frac{\partial L}{\partial y} \frac{\partial y}{\partial a_2} \frac{\partial a_2}{\partial z} \frac{\partial z}{\partial a_1} \frac{\partial a_1}{\partial w^{(1)}} \\
&= \eta(t-y)y'(a_2)w^{(2)}h'(a_1)x \\
&= \eta(t-y)y(1-y)w^{(2)}h'(a_1)x
\end{aligned}$$

Solution 6

(a)

Posterior distribution

$$p(\boldsymbol{w}|\mathcal{D}, \boldsymbol{m}_0, \alpha, \beta) \propto p(\boldsymbol{w}|\boldsymbol{m}_0, \alpha) \times p(D|\boldsymbol{w}, \beta)$$

where

$$p(\boldsymbol{w}|\boldsymbol{m}_0, \alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \alpha^{-1}I)$$

$$p(D|\boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(\boldsymbol{x}_n, \boldsymbol{w}), \beta^{-1})$$

Posterior predictive distribution

$$p(t|x, D, \beta, \boldsymbol{m}_0, \alpha) = \int p(t|x, \boldsymbol{w}) q(\boldsymbol{w}|D) d\boldsymbol{w} = \mathcal{N}(t|y(x, \boldsymbol{w}_{MAP}), \sigma^2(x))$$

where

$$y(x, \boldsymbol{w}) \simeq y(x, \boldsymbol{w}_{MAP}) + g^T(\boldsymbol{w} - \boldsymbol{w}_{MAP})$$

$$g = \nabla_{\boldsymbol{w}} y(x, \boldsymbol{w})|_{\boldsymbol{w}=\boldsymbol{w}_{MAP}}$$

$$\sigma^2(x) = \beta^{-1} + g^T A^{-1} g$$

(b)

Posterior distribution

$$p(\boldsymbol{w}|D,\alpha) = \int p(t|x,\boldsymbol{w})q(\boldsymbol{w}|D,\alpha)d\boldsymbol{w}$$

Prior predictive distribution

$$p(D|\beta,\boldsymbol{m}_0,\alpha) = \prod_{n=1}^{N} p(t|x,\alpha)$$

Solution 7

a) Please explain why the dual problem formulation is used to solve the SVM machine learning problem.

The dual formulation of SVM is favored because it simplifies the problem when data isn't linearly separable. It allows the use of kernels to handle higher-dimensional space efficiently. The dual form guarantees finding a global minimum, as it's a convex problem. It's computationally efficient since only support vectors (a subset of data) determine the decision boundary, leading to a sparse and faster-to-compute model.

b)

- **i) SVM vs Logistic Regression**: SVM focuses on the widest margin between classes, using hinge loss which ignores errors outside the margin. Logistic regression considers all data points with a logistic loss, predicting probabilities.
- **ii) v-SVM vs Least Squares Regression**: v-SVM uses ϵ-insensitive loss, ignoring errors within a certain range, which makes it robust to outliers. Least squares regression minimizes the sum of squared errors, sensitive to all deviations, making it less robust to outliers.

c) Neural networks use logistic activation functions for several reasons:

Neural networks use logistic activation functions to introduce non-linearity, enabling the network to learn complex patterns. Logistic functions are useful in output layers for binary classification since they produce probabilities (values between 0 and 1). They're continuously differentiable, a property needed for training algorithms like backpropagation. However, due to issues like gradient saturation, other functions like ReLU are now preferred in hidden layers.

d)

- **Differences**:
  - The logistic (sigmoid) function outputs values between 0 and 1, which is ideal for binary classifications.
  - The ReLU (Rectified Linear Unit) function outputs zero for negative inputs and raw input for positive inputs, which helps with the vanishing gradient problem and speeds up training.
  - The tanh function outputs values between -1 and 1, which centers the data, improving the learning for subsequent layers.
- **Usage**:
  - Logistic functions are often used in the output layer for binary classification problems.
  - ReLU is preferred in hidden layers due to its efficiency and effectiveness in deep networks.

    - Tanh is used when data centering is beneficial, but less common due to vanishing gradients.

e)

- The Jacobian matrix, representing first-order derivatives, is crucial for understanding the gradient of multivariate functions, helping in gradient descent optimization.
- The Hessian matrix, representing second-order derivatives, is used to find the curvature of the loss function, informing about the optimization landscape and guiding adjustments to the learning rate or direction.

f)

Exponential family distributions are common because they have convenient mathematical properties that allow for efficient estimation and inference, such as conjugate priors in Bayesian analysis.

Non-examples include the Cauchy distribution, which lacks a defined mean or variance, and the uniform distribution, which does not have a natural exponential form.

g)

Kullback-Leibler (KL) divergence measures how one probability distribution diverges from a second, expected probability distribution. It's useful for machine learning because it quantifies the difference between the learned model distribution and the true distribution of data.

- Example 1: In variational autoencoders (VAEs), KL divergence is used to regularize the encoder by penalizing divergences from the prior distribution, encouraging the latent space to approximate a standard normal distribution.
- Example 2: In natural language processing, KL divergence helps in comparing the similarity of word distribution in different text documents, aiding in tasks like topic modeling.

h)

Data augmentation techniques are a form of regularization for neural networks because they artificially increase the diversity of data available for training. By applying transformations like rotation, scaling, or cropping, they help the model generalize better, reduce overfitting, and improve robustness to variations in new, unseen data.

i)

- **Central Limit Theorem**: Many natural phenomena tend to follow a Gaussian distribution when they are the sum of many independent random variables.
- **Conjugacy**: Gaussian distributions are mathematically tractable, especially as conjugate priors in Bayesian inference, simplifying the computation of posterior distributions.
- **Continuity and Differentiability**: Gaussians are smooth and differentiable, which is beneficial for optimization algorithms that rely on gradient information.
- **Descriptive**: Gaussian distributions are defined by just two parameters (mean and variance), which can effectively capture the characteristics of many real-world datasets.

j)

it simplifies the computation by approximating the distribution around the peak (mode) with a Gaussian. This is effective when the peak is sharp, as it captures the main contribution to the integral, making it a practical approach for high-dimensional problems where exact computation is infeasible.

k)

Balance fit and complexity to prevent underfitting and overfitting. Cross-validation, information criteria (AIC, BIC), and regularization are used to evaluate and select models.

l)

features should be relevant, non-redundant, and have predictive power. Techniques like feature importance, correlation analysis, and domain knowledge can guide selection. For testing, samples should be representative but unseen during training to provide an unbiased evaluation.

An example is splitting a dataset into training and test sets, ensuring the test set includes a variety of examples across the feature space.

m)

The MAP model is often preferred over the ML model because it incorporates prior knowledge about the parameters through the prior distribution. This can lead to more reliable estimates, especially with small datasets. MAP can also mitigate overfitting by penalizing complex models, whereas ML estimates can overfit by focusing solely on the data likelihood.

Solution 8

(1)

Generative approaches to machine learning model how the data is generated, by learning the joint probability distribution $P(x, y)$. They can generate new data points and are powerful in unsupervised learning tasks. However, they can be complex and computationally intensive because they learn the full data distribution.

Discriminative approaches model the decision boundary between classes directly by learning the conditional probability $P(y|x)$. They often require less computation and provide better performance on classification tasks.

Example: Consider spam email filtering. A generative model would learn the distribution of words in both spam and non-spam emails to classify or even generate new email content. A discriminative model would learn the boundary that separates spam from non-spam based on features of the emails.

(2) **Analyzing the GAN Model**:

GAN combine both generative and discriminative models.

- The generative model in a GAN learns to produce data that's indistinguishable from real data, aiming to capture the data distribution.
- The discriminative model learns to distinguish real data from the fake data produced by the generator.