Outline I

1. Linear Regression Models
   - Basis Function
   - Least Squares
   - Bayesian Linear Regression
   - Evidence Approximation

2. Linear Models for Classification
   - Nonprobabilistic Approach
   - Probabilistic Approach

Polynomial Curve Fitting

Given a training set with

- N observations of x, $\mathbf{x} \equiv (x_1, \ldots, x_n)^\intercal$, and
- observations of target values of t, $\mathbf{t} \equiv (t_1, \ldots, t_n)^\intercal$

We shall fit the data using a polynomial function of the form

$$y(x, w) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M$$

by minimizing *error function*

$$E(w) = \frac{1}{2} \sum_{n=1}^{M} (y(x_n, w) - t_n)^2$$

▸ Poly Fit

Frequestist and Bayesian View

In a frequentist setting,

- w is considered to be a fixed parameter, whose value is determined by some form of "estimator", and
- error on this estimate are obtained by considering the distribution of possible observed data sets $\mathscr{D} = \{t_1, \dots, t_n\}$.

In a Bayesian setting,

- We assume a prior probability distribution $p(w)$ before observing the data.
- The effect of the observed data $\mathscr{D}$ is expressed through $p(\mathscr{D}|w)$, i.e., likelihood function.
- Bayes' theorem

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

  allows us to evaluate the uncertainty in w after we have observed $\mathscr{D}$ in the form of the posterior probability $p(w|\mathscr{D})$.

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Linear Basis Function Models

- We have input data $\mathscr{D}$ which consists of a set of D inputs $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_D\}$ and corresponding target values target values $\mathbf{t} = (t_1, \ldots, t_D)^T$.
- We assume that the target variable t is given by a deterministic function $y(\mathbf{x}, \mathbf{w})$ with additive Gaussian noise so that

$$t = y(\mathbf{x}, \mathbf{w}) + \varepsilon \text{ with } y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^\intercal \phi(\mathbf{x})$$

where ε is is a zero mean Gaussian random variable with precision β,

$$\mathbf{w} = (w_0, w_1, \ldots, w_{M-1})^\intercal$$

and

$$\phi(\mathbf{x}) = (\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \ldots, \phi_{M-1}(\mathbf{x}))^\intercal$$

are basis functions.

▸ Basis Funcs

Maximal Likelihood and Least Squares

Because the target variable t is given by a deterministic function $y(\mathbf{x}, \mathbf{w})$ plus a Gaussian noise ε with precision β[1]:

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}) \tag{1.1}$$

The log likelihood function is

$$\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_{\mathscr{D}}(\mathbf{w})$$

where

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

Theorem

Maximization of the likelihood function under a conditional Gaussian noise distribution for a linear model is equivalent to minimize a sum-of-squares error function given by $E_{\mathscr{D}}(\mathbf{w})$. The normal equations define $\mathbf{w}_{ML}$.

▸ Least Squares

[1] Gaussian noise implies that the conditional distribution of t given x is unimodal

Stochastic Gradient Descent

- When the training data set is very large or data is received in a stream, a direct solution using the normal equations may not be possible.
- An alternative approach is the *stochastic gradient descent* algorithm.
- The total error function

$$E_{\mathscr{D}}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\intercal \phi(x_n))^2 \equiv \sum_{n=1}^{N} E_n(\mathbf{w})$$

- In general, the stochastic gradient descent algorithm is applying

$$\mathbf{w}^{\tau+1} = \mathbf{w}^{\tau} - \eta \bigtriangledown_{\mathbf{w}} E_n$$

  where τ is the iteration number and η is a learning rate parameter.
- When the error function is the sum-of-squares function[1], then the algorithm is

$$\mathbf{w}^{\tau+1} = \mathbf{w}^{\tau} + \eta \left(t_n - \mathbf{w}^{(\tau)\intercal}\phi_n\right)\phi_n$$

▸ SGD

[1] For this type of total error function, the order of evaluation does not change the result

Regularization

One technique that is often used to control the over-fitting phenomenon is regularization which leads to a modified error function of the form

$$E_{\mathscr{D}}(\mathbf{w}) + \lambda E_{\mathbf{w}}(\mathbf{w})$$

Examples:

- $q = 1$ is know as the lasso in the statistics literature, and
- $q = 2$ corresponds to the quadratic regularizer.

▸ Regularization

Bias-Variance Decomposition

From a frequentist perspective, if consider a single input value $\mathbf{x}$, the expected squared loss can be decomposed as follows

$$\text{expected loss} = \text{bias}^2 + \text{variance} + \text{noise}$$

It is of limited practical value, because

- the bias-variance decomposition is based on averages with respect to ensembles of data sets $\mathbb{E}_{\mathscr{D}}$, whereas in practice we have only the single observed data set.
- If we had a large number of independent training sets of a given size, we would be better off combining them into a single large training set, which of course would reduce the level of over-fitting for a given model complexity.

▸ Bias-Variance Decomp

Bayesian Linear Regression

Suppose the noise precision parameter β is known. The likelihood function is

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

The corresponding conjugate prior $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0},\alpha^{-1}\mathbf{I})$. The posterior distribution is

$$p(\mathbf{w}|\mathbf{t},\mathbf{X}) \propto p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta)p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N)$$

where[1]

$$\mathbf{m}_N = \beta\mathbf{S}_N\mathbf{\Phi}^\intercal\mathbf{t} \text{ and } \mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^\intercal\mathbf{\Phi}$$

The log of the posterior distribution takes the form

$$\ln p(\mathbf{w}|\mathbf{t},\mathbf{X}) = -\frac{\beta}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\intercal\phi(\mathbf{x}_n))^2 - \frac{\alpha}{2}\mathbf{w}^\intercal\mathbf{w} + \text{const.}$$

▸ Bayesian Reg

[1] $\mathbf{x}$ will always appear in the set of conditioning variables. We may drop the explicit $\mathbf{x}$ from future expressions for simplicity. General prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0,\mathbf{S}_0)$ gives $\mathbf{m}_N = \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^\intercal\mathbf{t}\right)$ and $\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^\intercal\mathbf{\Phi}$

Bayesian Linear Regression

Theorem

Maximization of the posterior distribution with respect to **w** is therefore equivalent to the minimization of the sum-of-squares error function with the addition of a quadratic regularization term with $\lambda = \alpha/\beta$.

- Maximum likelihood provides a point estimate of **w**;
- Bayesian method provides a distribution of w, which gives a predictive distribution.

To predict t for new values of **x**. The predictive distribution is

$$p(t|\mathbf{t},\mathbf{x},\alpha,\beta) = \int p(t|\mathbf{w},\mathbf{x},\beta)p(\mathbf{w}|\mathbf{t},\mathbf{x},\alpha,\beta)d\mathbf{w}$$

It can be shown that

$$p(t|\mathbf{t},\mathbf{x},\alpha,\beta) = \mathcal{N}\left(t|\mathbf{m}_N^\intercal\phi(\mathbf{x}),\sigma_N^2(\mathbf{x})\right)$$

where $\sigma_N^2(\mathbf{x}) = \beta^{-1} + \phi(\mathbf{x})^\intercal \mathbf{S}_N \phi(\mathbf{x})$.

▸ Bayesian Prediction

- The first term represents the noise on the data
- The second term reflects the uncertainty associated with the parameters w.
- Because the noise process and the distribution of **w** are independent Gaussians, their variances are additive.

Bayesian Model Comparison

Compare a set of model $\mathscr{M}_i$. Given a training set $\mathscr{D}$, we wish to evaluate

$$p(\mathscr{M}_i|\mathscr{D}) \propto p(\mathscr{M}_i)p(\mathscr{D}|\mathscr{M}_i)$$

$p(\mathscr{M}_i)$ allows us to express a preference for different models. By simply assuming equal prior, $p(\mathscr{D}|\mathscr{M}_i)$ is model evidence or marginal likelihood as the parameters are marginalized out.

$$p(\mathbf{w}|\mathscr{D},\mathscr{M}_i) = \frac{p(\mathscr{D}|\mathbf{w},\mathscr{M}_i)p(\mathbf{w}|\mathscr{M}_i)}{p(\mathscr{D}|\mathscr{M}_i)} \Rightarrow p(\mathscr{D}|\mathscr{M}_i) = \int p(\mathscr{D}|\mathbf{w},\mathscr{M}_i)p(\mathbf{w}|\mathscr{M}_i)d\mathbf{w}$$

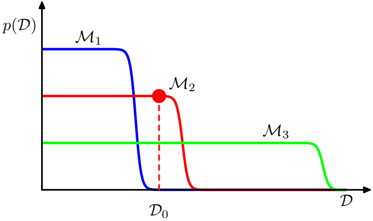
$$\ln p(\mathscr{D}) = \ln\left(\int p(\mathscr{D}|\mathbf{w})p(\mathbf{w})d\mathbf{w}\right) \approx \ln\left(p(\mathscr{D}|\mathbf{w}_{\text{MAP}})\int_{\text{posterior support}} \frac{1}{\Delta\mathbf{w}_{\text{prior}}}d\mathbf{w}\right)$$

$$\approx \ln p(\mathscr{D}|\mathbf{w}_{\text{MAP}}) + M\ln\frac{\Delta\mathbf{w}_{\text{posterior}}}{\Delta\mathbf{w}_{\text{prior}}}$$

- model complexity ↑ ⇒ first term ↓, because complex model fits data better.
- model complexity ↑ ⇒ second term[1] ↑ due to M

[1] All parameters have the same ratio of $\Delta\mathbf{w}_{\text{posterior}}/\Delta\mathbf{w}_{\text{prior}}$

Bayesian Model Comparison



- The Bayesian framework avoids the problem of over-fitting and allows models to be compared on the basis of the training data alone.
- A simple model (such as a first order polynomial) has little variability and so will generate data sets that are fairly similar to each other. Its distribution $p(\mathscr{D})$ is therefore confined to a relatively small region of the horizontal axis.
- A complex model (such as a ninth order polynomial) generates a great variety of different data sets, so its distribution $p(\mathscr{D})$ is spread over a large region of the space of data sets
- The model evidence can be sensitive to many aspects of the prior.

Predictive Distribution

In practice, we are interested in making predictions of t for new values of $\mathbf{x}$. This requires that we evaluate the predictive distribution defined by

$$p(t|\mathbf{t}, \mathbf{x}, \alpha, \beta) = \int p(t|\mathbf{w}, \mathbf{x}, \beta) p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \alpha, \beta) d\mathbf{w}$$

It can be shown that

$$p(t|\mathbf{t}, \mathbf{x}, \alpha, \beta) = \mathcal{N}\left(t|\mathbf{m}_N^\intercal \phi(\mathbf{x}), \sigma_N^2(\mathbf{x})\right)$$

where $\sigma_N^2(\mathbf{x}) = \beta^{-1} + \phi(\mathbf{x})^\intercal \mathbf{S}_N \phi(\mathbf{x})$.

- The first term represents the noise on the data whereas
- The second term reflects the uncertainty associated with the parameters w.
- Because the noise process and the distribution of $\mathbf{w}$ are independent Gaussians, their variances are additive.

▸ Bayesian Prediction

Evidence Approximation

In a fully Bayesian treatment, the predictive distribution is

$$p(t|\mathbf{t}) = \int\int\int p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)p(\alpha,\beta|\mathbf{t})d\mathbf{w}d\alpha d\beta$$

- $p(t|\mathbf{w},\beta)$: target t is determined by $\mathbf{w}$ and Gaussian noise;
- $p(\mathbf{w}|\mathbf{t},\alpha,\beta)$: posterior distribution of $\mathbf{w}$;
- $p(\alpha,\beta|\mathbf{t})$: posterior distribution of hyper-parameters.

An evidence approximation (if $p(\alpha,\beta|\mathbf{t})$ is sharply peaked around $\hat{\alpha},\hat{\beta}$)

$$p(t|\mathbf{t}) \approx p(t|\mathbf{t},\hat{\alpha},\hat{\beta}) = \int\int p(t|\mathbf{w},\hat{\beta})p(\mathbf{w}|\mathbf{t},\hat{\alpha},\hat{\beta})d\mathbf{w}$$

Note that $p(\alpha,\beta|\mathbf{t}) \propto p(\mathbf{t}|\alpha,\beta)p(\alpha,\beta)$. If prior $p(\alpha,\beta)$ is flat, $\hat{\alpha},\hat{\beta}$ can be obtained by maximizing the marginal likelihood $p(\mathbf{t}|\alpha,\beta)$, where

$$p(\mathbf{t}|\alpha,\beta) = \int p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)d\mathbf{w}$$

- $p(\mathbf{t}|\mathbf{w},\beta)$: the likelihood function;
- $p(\mathbf{w}|\alpha)$: prior distribution of $\mathbf{w}$.

Empirical Bayes

Algorithm

1. Initialization: $k = 0$, $\alpha^0 = \alpha_0$ and $\beta^0 = \beta_0$
2. Find eigenvalues λ_i $i = 0, \ldots, M-1$ such that

$$\left(\beta \Phi^T \Phi\right) \mathbf{u}_i = \lambda_i \mathbf{u}_i$$

3. Let

$$\gamma^k = \sum_{i=0}^{M-1} \frac{\lambda_i}{\alpha^k + \lambda_i}, \ \alpha^{k+1} = \frac{\gamma^k}{\mathbf{w}_{mean}^T \mathbf{w}_{mean}} \text{ and } \frac{1}{\beta^{k+1}} = \frac{1}{N-\gamma} \sum_{i=1}^{N} \left(t_i - \mathbf{w}_{mean}^T \phi(x_i)\right)^2$$

   where $\mathbf{w}_{mean} = \mathbf{m}_N = \beta \mathbf{S}_N \mathbf{\Phi}^T \mathbf{t}$.

4. If $|\alpha^{k+1} - \alpha^k| + |\beta^{k+1} - \beta^k| <$ threshold, then return α, β, else $k = k+1$ and go to step 2.

▸ Empirical Bayes

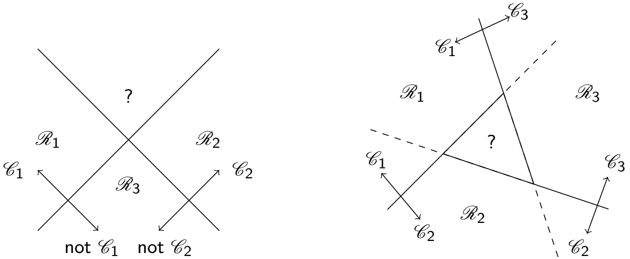
Introduction

- The goal in classification is to assign D-dimension $\mathbf{x}$ to one of K classes $\mathscr{C}$.
- A target vector $\mathbf{t} = (0,1,0,0,0)^\intercal$ indicates a pattern from class 2 out of 5 classes and we can interpret the value of t_k as the probability that the class is $\mathscr{C}_k$.
- Nonprobabilistic approach constructs a **discriminant** function that directly assigns each vector $\mathbf{x}$ to a specific class.
- Probabilistic approach models the conditional probability distribution $p(\mathscr{C}_k|\mathbf{x})$ in an **inference** stage, and then subsequently uses this distribution to make optimal decisions.

▸ Training Data

Discriminant Function

A linear function $y(\mathbf{x}) = \mathbf{w}^\mathsf{T}\mathbf{x} + w_0$ can assign $\mathbf{x}$ to class $\mathscr{C}_1$ if $y(\mathbf{x}) \geq 0$ and to class $\mathscr{C}_2$ otherwise. However, for multiple classes



(a) one-versus-the-rest classifier

(b) one-versus-one classifier

- Figure (a) uses $K-1$ classifier each of which solves a two-class problem.
- Figure (b) uses $K(K-1)/2$ classifier and one for every possible pair of classes.
- Both run into the problem of ambiguous regions.

Multiple K Classes

A single K-class discriminant: $y_k(\mathbf{x}) = \mathbf{w}_k^\intercal \mathbf{x} + w_{k0}$. $\mathbf{x}$ is assigned to class $\mathscr{C}_k$ if $y_k(\mathbf{x}) > y_j(\mathbf{x})$ for all $j \neq k$, i.e. class $k^* = \mathrm{argmax}_k \{y_k(\mathbf{x}) : k = 1, \ldots, K\}$.

Least Squares for Classification

Consider a training data set $\{\mathbf{x}_i, \mathbf{t}_i\}$ where $i = 1, \ldots, N$. The least squares approach is to find $\mathbf{w}_k$, $k = 1, \ldots, K$ such that the sum-of-squares error between $\mathbf{y}(\mathbf{x_i}) = (y_1(\mathbf{x}), \ldots, y_K(\mathbf{x}))$ and $\mathbf{t}_i$ is minimal.

Remark

The failure of least squares should not surprise us when we recall that it corresponds to maximum likelihood under the assumption of a Gaussian conditional distribution, whereas binary target vectors clearly have a distribution that is far from Gaussian.

▸ Least Squares

Linear Discriminant

Linear Classification

A linear classification model reduce the D dimension input $\mathbf{x}$ down to one dimension using $y = \mathbf{w}^{\mathsf{T}}\mathbf{x}$

consider a two-class problem in which there are N_1 points of class $\mathscr{C}_1$ and N_2 points of class $\mathscr{C}_2$, so that the mean vectors of two classes:

$$\mathbf{m}_1 = \frac{1}{N_1}\sum_{n\in\mathscr{C}_1}\mathbf{x}_n \text{ and } \mathbf{m}_2 = \frac{1}{N_2}\sum_{n\in\mathscr{C}_2}\mathbf{x}_n$$

We want to choose $\mathbf{w}$ to maximize $\mathbf{w}^{\mathsf{T}}(\mathbf{m}_1 - \mathbf{m}_2) \equiv m_1 - m_2$, with scaler $\sum_i w_i^2 = 1$.

Fisher's Method

Fisher's idea is to maximize a function that will give a large separation between the projected class means while giving a small variance within each class, thereby minimizing the class overlap.

$$J(\mathbf{w}) = \frac{(m_1 - m_2)^2}{\sum_{k=1}^{2}\sum_{n\in\mathscr{C}_k}(y_n - m_k)^2} = \frac{\text{between class variance}}{\text{within class variance}}$$

Probabilistic Approach

- Generative models consider

$$\underbrace{p(\mathscr{C}_k|\mathbf{x})}_{\text{posterior probabilities}} = \underbrace{p(\mathbf{x}|\mathscr{C}_k)}_{\text{class-conditional densities}} \cdot \underbrace{p(\mathscr{C}_k)}_{\text{class priors}} \equiv f_k(a_k).$$

  When class-conditional densities are in exponential family, $f_k = \text{softmax}_k$ (or $\sigma(\cdot)$ logistic sigmoid function if $K = 2$); a_k is a linear function of $\mathbf{x}$.

- Generative models use maximum likelihood solution to estimate parameters in $p(\mathbf{x}|\mathscr{C}_k)$ and $p(\mathscr{C}_k)$.
- Discriminative models consider $p(\mathscr{C}_k|\mathbf{x}) = f_k(a_k)$ where $a_k = \mathbf{w}^\intercal \phi$ directly with nonlinear transfer ϕ.
- $f_k = \sigma(\cdot)$ is logistic regression and $f_k = \Phi(\cdot)$ is probit regression.

▸ Logistic

- In discriminative models, $\mathbf{w}$ can be estimated by least squares or Bayesian approach.

▸ Multiple Methods

Laplace Approximation, AIC and BIC

Laplace Approximation is to find a Gaussian approximation to a probability density $p(z)$

$$p(z) = \frac{1}{Z} f(z) \text{ where } Z = \int f(z) dz$$

$$p(z) \sim \mathcal{N}(z|z_0, A^{-1}) \text{ where } z_0 \text{ is a mode of } p(z)$$

$f'(z_0) = 0$ and $A = -\nabla\nabla \ln f(z_0)$.

- AIC: $\ln p(\mathscr{D}|\mathbf{w}_{\text{ML}}) - M$, where $p(\mathscr{D}|\mathbf{w}_{\text{ML}})$ is the best-fit log likelihood.
- BIC: recall Bayesian model evidence

$$\ln p(\mathscr{D}) \approx \ln p(\mathscr{D}|\mathbf{w}_{\text{MAP}}) + \underbrace{\ln p(\mathbf{w}_{\text{ML}}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|A|}_{\text{Occam factor}}$$

$$\approx \ln p(\mathscr{D}|\mathbf{w}_{\text{MAP}}) - \frac{1}{2} M \ln N$$