Parameter Estimation for Generative Models

Prof. Ziping Zhao

School of Information Science and Technology
ShanghaiTech University, Shanghai, China

CS182: Introduction to Machine Learning (Fall 2021)
http://cs182.sist.shanghaitech.edu.cn

Outline

Outline

Different Approaches to Supervised Learning

- Three major approaches corresponding to different model assumptions:
  - Parametric approach
  - Nonparametric approach
  - Semiparametric approach
- Current topic: parametric approach

Parametric Approach

- Assumption: data follows a distribution that obeys a parametric model, e.g., Gaussian.
- Advantage of the parametric approach: the model is fully specified by a small number of parameters θ as sufficient statistics of the distribution
- The sample $\mathcal{X} \in \{\mathbf{x}^{(\ell)}\}$ is assumed to be drawn (usually i.i.d.) from the underlying distribution, i.e., $\mathbf{x}^{(\ell)} \sim p(\mathbf{x})$
- The number of parameters $\dim(\theta)$ is independent of the sample size $|\mathcal{X}|$.
- Parameter estimation: assuming some parametric form (as a kind of inductive bias) for $p(\mathbf{x} \mid \theta)$, θ is estimated using $\mathcal{X}$ (density estimation).
- The estimated model/distribution is then used to make a decision
- Two approaches to parameter estimation:
  - Maximum likelihood estimation: θ is a fixed point (point estimation)
  - Bayesian estimation: θ is a random variable whose prior uncertainty (represented as prior distribution) can be incorporated.

Parametric Approach to Classification

- Recall Bayes' rule for classification:

$$P(C_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_i)P(C_i)}{\sum_{k=1}^{K} p(\mathbf{x} \mid C_k)P(C_k)}$$

$$\text{Choose } C_i \text{ if } P(C_i \mid \mathbf{x}) = \max_k P(C_k \mid \mathbf{x})$$

- To compute $P(C_i \mid x)$ for classification, $p(\mathbf{x} \mid C_i)$ and $P(C_i)$ have to be estimated from the sample $\mathcal{X}$.
- A classifier thus created is often called a generative classifier (as opposed to a discriminative classifier to be studied later in the course), since it specifies how to generate the data based on $p(\mathbf{x} \mid C_i)$ and $P(C_i)$.

Maximum Likelihood Estimation

- Maximum likelihood estimation (MLE) seeks to find θ that makes sampling $\mathcal{X}$ from $p(\mathbf{x} \mid \theta)$ as likely as possible by maximizing the likelihood of θ given the sample $\mathcal{X} = \{\mathbf{x}^{(\ell)}\}_{\ell=1}^{N}$.
- Likelihood of θ given $\mathcal{X}$ (with the i.i.d. assumption):

$$L(\theta \mid \mathcal{X}) = p(\mathcal{X} \mid \theta) = \prod_{\ell=1}^{N} p(\mathbf{x}^{(\ell)} \mid \theta)$$

- Log likelihood (mainly for computational simplification):

$$\mathcal{L}(\theta \mid \mathcal{X}) = \log L(\theta \mid \mathcal{X}) = \sum_{\ell=1}^{N} \log p(\mathbf{x}^{(\ell)} \mid \theta)$$

- Maximum likelihood estimate:

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta \mid \mathcal{X})$$

Outline

Outline

Example: Bernoulli

- Discrete random variable x with two possible values, $x \in \{0, 1\}$
- E.g., use $P(x = 1)$ to represent $P(C_1)$ (and hence $P(x = 0) = 1 - P(x = 1)$ represents $P(C_2)$).
- Probability mass function (with parameter $\theta = p = P(x = 1)$):

$$P(x \mid p) = p^x (1 - p)^{1-x}$$

- Log likelihood:

$$\mathcal{L}(p \mid \mathcal{X}) = \sum_{\ell=1}^{N} \left[x^{(\ell)} \log p + (1 - x^{(\ell)}) \log(1 - p) \right]$$

- ML estimate:

$$\hat{p} = \frac{1}{N} \sum_{\ell=1}^{N} x^{(\ell)} \quad \text{(sample mean)}$$

Example: Bernoulli (2)

- Note that the estimate $\hat{p}$ is a function of the sample and is another random variable
- We can discuss the distribution of $\hat{p}(\mathcal{X}_i)$ given different $\mathcal{X}_i$'s sampled from the same $p(x)$
  - the variance of the distribution of $\hat{p}(\mathcal{X}_i)$ is expected to decrease as N increases; as the samples get bigger, they (and hence their averages) get more similar.

Example: Multinomial

- Discrete random variable x with $K \geq 2$ possible values, e.g., for K classes.
- Generalization of Bernoulli distribution.
- Indicator variables $x_1, \ldots, x_K$:

$$x_i = \begin{cases} 1 & \text{if outcome is state } i \\ 0 & \text{if outcome is not state } i \end{cases}$$

  with $\sum_{i=1}^{K} x_i = 1$.
- Probability mass function (with parameters $\theta = (p_1, \ldots, p_K)^T$):

$$P(x \mid \theta) = P(x_1, \ldots, x_K \mid p_1, \ldots, p_K) = \prod_{i=1}^{K} p_i^{x_i}$$

  with constraint

$$\sum_{i=1}^{K} p_i = 1$$

Example: Multinomial (2)

- Log likelihood:

$$\mathcal{L}(p_1, \ldots, p_K \mid \mathcal{X}) = \sum_{\ell=1}^{N} \sum_{i=1}^{K} x_i^{(\ell)} \log p_i$$

- Constrained optimization problem (with equality constraint $\sum_{i=1}^{K} p_i = 1$) which can be solved using the method of Lagrange multipliers.
- ML estimates:

$$\hat{p}_i = \frac{1}{N} \sum_{\ell=1}^{N} x_i^{(\ell)}$$

  So $\hat{\theta} = (\hat{p}_1, \ldots, \hat{p}_K)^T$ is also the sample mean.
- Another estimation method: view it as K separate Bernoulli experiments

Example: Normal

- Continuous random variable x following univariate normal (Gaussian) distribution $\mathcal{N}(\mu, \sigma^2)$ with mean μ and variance σ^2.
- Probability density function (with parameters $\theta = (\mu, \sigma)^T$):

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

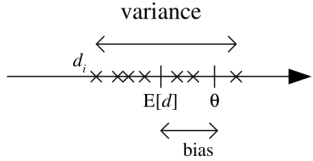
- Log likelihood:

$$\mathcal{L}(\mu, \sigma \mid \mathcal{X}) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{1}{2\sigma^2}\sum_{\ell=1}^{N}(x^{(\ell)} - \mu)^2$$

- ML estimates:

$$\hat{\mu} = \frac{1}{N}\sum_{\ell=1}^{N} x^{(\ell)} \qquad \hat{\sigma}^2 = \frac{1}{N}\sum_{\ell=1}^{N}(x^{(\ell)} - \hat{\mu})^2$$

Bias and Variance



- Parameter to be estimated: θ
- Estimator based on sample $\mathcal{X}_i$: $d_i = d(\mathcal{X}_i)$
- Bias: measures how much the expected value of the estimate $\mathbb{E}[d] = \mathbb{E}[d(\mathcal{X})]$ deviates from the correct value θ.

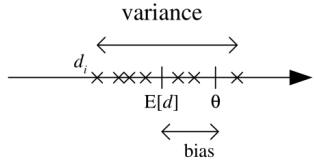
$$b_\theta[d] = \mathbb{E}[d] - \theta$$

Variance: measures how much, on average, the estimate d varies from the expected value $\mathbb{E}[d]$.

$$Var(d) = \mathbb{E}\left[(d - \mathbb{E}[d])^2\right]$$

- We would like both to be small.

Bias and Variance (2)

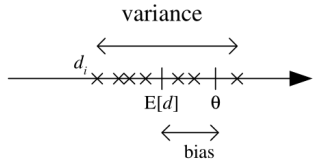


- Mean squared error (MSE) of estimator d (measures how much d is different from θ):

$$\begin{aligned} r(d,\theta) &= \mathbb{E}\left[(d-\theta)^2\right] \\ &= \mathbb{E}\left[(d-\mathbb{E}[d]+\mathbb{E}[d]-\theta)^2\right] \\ &= \dots \\ &= (\mathbb{E}[d]-\theta)^2 + \mathbb{E}\left[(d-\mathbb{E}[d])^2\right] \\ &= (b_\theta[d])^2 + Var(d) = \text{bias}^2 + \text{variance} \end{aligned}$$

- Bias-variance tradeoff

Bias and Variance (3)

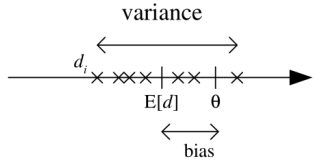


- If $b_\theta[d] = \mathbb{E}[d] - \theta = 0$ for all θ, d is an unbiased estimator of θ.
  - $x^{(\ell)}$ is from some density with mean μ, then the sample mean $\hat{\mu} = \frac{1}{N}\sum_{\ell=1}^{N} x^{(\ell)}$ is an unbiased estimator of the mean μ, since

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}[\frac{1}{N}\sum_{\ell=1}^{N} x^{(\ell)}] = \frac{1}{N}\sum_{\ell=1}^{N}\mathbb{E}[x^{(\ell)}] = \frac{1}{N}N\mu = \mu$$

  - $N \to \infty$, $\hat{\mu} \to \mu$

Bias and Variance (4)



- sample mean $\hat{\mu}$ is a consistent estimator of μ, since $Var(\hat{\mu}) \to 0$ as $N \to \infty$
  - $x^{(\ell)}$ is from some density with variance σ^2

$$Var(\hat{\mu}) = Var[\frac{1}{N}\sum_{\ell=1}^{N} x^{(\ell)}] = \frac{1}{N^2}\sum_{\ell=1}^{N} Var[x^{(\ell)}] = \frac{1}{N^2} N\sigma^2 = \frac{\sigma^2}{N}$$

  - as N gets larger, $\hat{\mu}$ deviates less from μ

Outline

Bayesian Estimation

- Unlike MLE which treats θ as a fixed (but unknown) point, the Bayesian estimation approach treats it as a random variable with prior density $p(\theta)$ modeling our a prior uncertainty about θ.
- Posterior density of θ (i.e., obtaining uncertainty about θ after observing the sample $\mathcal{X}$ by combining the likelihood density $p(\mathcal{X} \mid \theta)$):

$$p(\theta \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \theta)p(\theta)}{p(\mathcal{X})} = \frac{p(\mathcal{X} \mid \theta)p(\theta)}{\int p(\mathcal{X} \mid \theta')p(\theta')d\theta'}$$

- Full Bayesian estimation approach (through marginalization over θ):
  - Estimation of the density at any x (i.e., the propability that any sample occurs):

$$p(x \mid \mathcal{X}) = \int p(x, \theta \mid \mathcal{X})d\theta = \int p(x \mid \theta, \mathcal{X})p(\theta \mid \mathcal{X})d\theta = \int p(x \mid \theta)p(\theta \mid \mathcal{X})d\theta$$

    taking an average over predictions using all θ, weighted by their probabilities
  - Prediction (e.g., regression) based on a function $g(x \mid \theta)$:

$$y = \int g(x \mid \theta)p(\theta \mid \mathcal{X})d\theta$$

Computational Considerations

- Evaluating the integrals (for marginalization over θ) may be computationally difficult, except in cases where the posterior has a nice form.
- So the full Bayesian estimation approach is sometimes replaced by other methods for computational considerations.
- Maximum a posteriori (MAP) estimation - mode of the posterior density:

$$\theta_{\text{MAP}} = \arg\max_{\theta} p(\theta \mid \mathcal{X}) = \arg\max_{\theta} p(\mathcal{X} \mid \theta) p(\theta)$$

$$p(x \mid \mathcal{X}) = p(x \mid \theta_{\text{MAP}}) \qquad y_{\text{MAP}} = g(x \mid \theta_{\text{MAP}})$$

  Maximum likelihood (ML) estimation – MAP with flat prior:

$$\theta_{\text{ML}} = \arg\max_{\theta} p(\mathcal{X} \mid \theta)$$

- Bayes' estimator – expectation w.r.t. posterior density:

$$\theta_{\text{Bayes}} = \mathbb{E}[\theta \mid \mathcal{X}] = \int \theta p(\theta \mid \mathcal{X}) d\theta$$

Bayes' estimator

- Bayes' estimator – expectation w.r.t. posterior density:

$$\theta_{\text{Bayes}} = \mathbb{E}[\theta \mid \mathcal{X}] = \int \theta p(\theta \mid \mathcal{X}) d\theta$$

- The reason for taking the expected value is that the best estimate of a random variable is its mean.
- Let us say θ is the variable we want to predict with $\mathbb{E}[\theta] = \mu$. For any estimate c (a constant value), we have

$$\begin{aligned}\mathbb{E}[(\theta - c)^2] &= \mathbb{E}[(\theta - \mu + \mu - c)^2] \\ &= \mathbb{E}[(\theta - \mu)^2] + (\mu - c)^2\end{aligned}$$

  which is minimum if c is taken as μ.
- In the case of a normal density, the mode is the expectation.
  - If $p(\theta \mid \mathcal{X})$ is normal, then $\theta_{\text{MAP}} = \theta_{\text{Bayes}}$.

Example

- Bayesian estimation with known μ_0, σ_0 and σ:

$$x^{(\ell)} \sim \mathcal{N}(\theta, \sigma^2),\ \theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

- MLE:

$$\theta_{\mathrm{ML}} = \frac{1}{N}\sum_{\ell=1}^{N} x^{(\ell)} = m \quad (\text{sample mean})$$

- The $p(\theta \mid \mathcal{X})$ is normal. MAP and Bayes' estimator:

$$\theta_{\mathrm{MAP}} = \theta_{\mathrm{Bayes}} = \mathbb{E}[\theta \mid \mathcal{X}] = \frac{1/\sigma_0^2}{N/\sigma^2 + 1/\sigma_0^2}\mu_0 + \frac{N/\sigma^2}{N/\sigma^2 + 1/\sigma_0^2} m$$

  weighted average of the sample mean m and the prior mean μ_0, with weights being inversely proportional to their variances
  - As N increases, the Bayes' estimator gets closer to the sample average, using more the information provided by the sample.
  - When σ_0^2 is small (we have little prior uncertainty regarding the correct value of θ), or when N is small, our prior guess μ_0 has a higher effect.

Bayesian estimation

- Both MAP and Bayes' estimators reduce the whole posterior density to a single point and lose information unless the posterior is unimodal and makes a narrow peak around these points.
- With computation getting cheaper, we can use a Monte Carlo approach that generates samples from the posterior density.
- There also are many approximation methods one can use to evaluate the full integral in the full Bayesian estimation approach.

Outline

Classification with Discriminant Functions

- In Bayes' decision rule for classification, the discriminant function for of class C_i is

$$p(x \mid C_i)P(C_i) \quad \text{or} \quad \log\left[p(x \mid C_i)P(C_i)\right]$$

- Assume Gaussian density for each class:

$$p(x \mid C_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right]$$

- Discriminant functions:

$$\begin{aligned} g_i(x) &= \log\left[p(x \mid C_i)P(C_i)\right] \\ &= \log p(x \mid C_i) + \log P(C_i) \\ &= -\frac{1}{2}\log 2\pi - \log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log P(C_i) \end{aligned}$$

Discriminant Functions based on ML Estimates

- Samples $\mathcal{X} = \{(x^{(\ell)}, \mathbf{y}^{(\ell)})\}_{\ell=1}^{N}$ where

$$x^{(\ell)} \in \mathbb{R} \qquad y_i^{(\ell)} = \begin{cases} 1 & \text{if } x^{(\ell)} \text{ belongs to } C_i \\ 0 & \text{if } x^{(\ell)} \text{ belongs to } C_k, k \neq i \end{cases}$$

- ML estimates:

$$\hat{P}(C_i) = \frac{1}{N} \sum_{\ell=1}^{N} y_i^{(\ell)} \text{ (multinomial density)}$$

$$m_i = \frac{\sum_{\ell=1}^{N} x^{(\ell)} y_i^{(\ell)}}{\sum_{\ell=1}^{N} y_i^{(\ell)}} \qquad s_i^2 = \frac{\sum_{\ell=1}^{N} (x^{(\ell)} - m_i)^2 y_i^{(\ell)}}{\sum_{\ell=1}^{N} y_i^{(\ell)}} \text{ (normal density)}$$

- Discriminant functions (with constant term $-\frac{1}{2} \log 2\pi$ dropped):

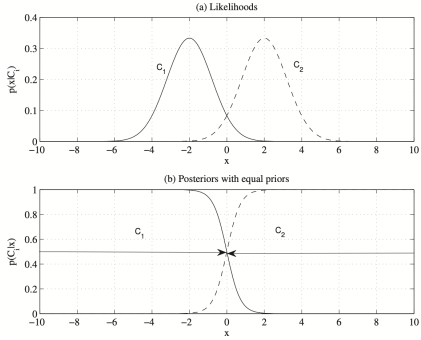
$$g_i(x) = -\log s_i - \frac{(x - m_i)^2}{2s_i^2} + \log \hat{P}(C_i)$$

Special Case: Equal Priors and Variances

- Simplified discriminant functions: $g_i(x) = -(x - m_i)^2$ (the quadratic term x^2 can be reduced since it is common in all discriminants, leading to a linear discriminant)
- Classification rule (nearest mean classifier):

$$\text{Choose } C_i \text{ if } |x - m_i| = \min_k |x - m_k|$$

- Likelihood functions and posterior densities:

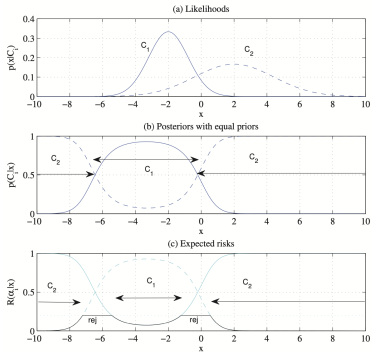


Special Case: Equal Priors but Different Variances

- Simplified discriminant functions:

$$g_i(x) = -\log s_i - \frac{(x - m_i)^2}{2s_i^2}$$

- Likelihood functions, posterior densities, and expected risks (for reject with $\lambda = 0.2$):



Special Case: Different Priors

- If the priors are different, this has the effect of moving the threshold of decision toward the mean of the less likely class.

Discriminant Functions based on Bayesian Estimates

- We have used the maximum likelihood estimators for the parameters.
- If we have some prior information about them, for example, for the means, we can use a Bayesian estimate of $p(x \mid C_i)$ with prior on μ_i.

Remarks

Gaussianity

- When x is continuous, $p(x \mid C_i)$ cannot be immediately assumed as Gaussian densities.

Generative vs. Discriminative Classifiers

- This is the generative approach (likelihood-based approach) to classification where we use data to estimate the densities separately, calculate posterior densities using Bayes' rule, and then get the discriminant.
- We will discuss the discriminative approach (discriminant-based approach) in later lectures where we bypass the estimation of densities and directly estimate the discriminants.

Outline

Additive Parametric Model

- Functional relationship between output (dependent variable) and input (independent variable(s)) expressed in an additive form:

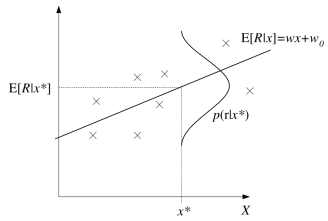
$$r = f(x) + \epsilon$$

  where $f(x)$ is an unknown but deterministic function and ϵ is random noise independent of the input.
- Parametric modeling:
  - $f(x) \approx$ estimator $g(x \mid \theta)$ (defined up to a set of parameters θ)
  - $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Maximum Likelihood Estimation

- Conditional probability of output given input:

$$p(r \mid x) \sim \mathcal{N}(g(x \mid \theta), \sigma^2)$$



- Joint density:

$$p(x, r) = p(r \mid x)p(x)$$

where $p(x)$ is the input density.

Maximum Likelihood Estimation (2)

- Log likelihood of θ given i.i.d. sample $\mathcal{X} = \{(x^{(\ell)}, r^{(\ell)})\}_{\ell=1}^{N}$:

$$\mathcal{L}(\theta \mid \mathcal{X}) = \log \prod_{\ell=1}^{N} p(x^{(\ell)}, r^{(\ell)}) = \sum_{\ell=1}^{N} \log p(r^{(\ell)} \mid x^{(\ell)}) + \sum_{\ell=1}^{N} \log p(x^{(\ell)})$$

$$= \sum_{\ell=1}^{N} \log p(r^{(\ell)} \mid x^{(\ell)}) + \text{const.} = \cdots = -\frac{1}{2\sigma^2} \sum_{\ell=1}^{N} \left[r^{(\ell)} - g(x^{(\ell)} \mid \theta)\right]^2 + \text{const.}$$

- Maximizing $\mathcal{L}(\theta \mid X)$ is equivalent to minimizing the following error function:

$$E(\theta \mid X) = \frac{1}{2} \sum_{\ell=1}^{N} \left[r^{(\ell)} - g(x^{(\ell)} \mid \theta)\right]^2$$

- So the ML estimate of θ under Gaussian error assumption is also the least squares estimate.

Linear Regression

- Linear regression function:

$$g(x^{(\ell)} \mid w_0, w_1) = w_1 x^{(\ell)} + w_0$$

- Error function:

$$E(w_0, w_1 \mid \mathcal{X}) = \frac{1}{2} \sum_{\ell=1}^{N} (r^{(\ell)} - w_1 x^{(\ell)} - w_0)^2$$

- Setting the derivatives of $E(w_0, w_1 \mid X)$ w.r.t. w_0, w_1 to 0 gives

$$\sum_{\ell=1}^{N} r^{(\ell)} = N w_0 + w_1 \sum_{\ell=1}^{N} x^{(\ell)}$$

$$\sum_{\ell=1}^{N} r^{(\ell)} x^{(\ell)} = w_0 \sum_{\ell=1}^{N} x^{(\ell)} + w_1 \sum_{\ell=1}^{N} \left(x^{(\ell)}\right)^2$$

Linear Regression (2)

- Linear system in vector-matrix form:

$$\mathbf{A}\mathbf{w} = \mathbf{y}$$

where

$$\mathbf{A} = \begin{bmatrix} N & \sum_\ell x^{(\ell)} \\ \sum_\ell x^{(\ell)} & \sum_\ell (x^{(\ell)})^2 \end{bmatrix}$$

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} \sum_\ell r^{(\ell)} \\ \sum_\ell r^{(\ell)} x^{(\ell)} \end{bmatrix}$$

- Least squares estimate (assuming that $\mathbf{A}$ is invertible):

$$\hat{\mathbf{w}} = \mathbf{A}^{-1}\mathbf{y}$$

Polynomial Regression

- Polynomial regression function of order k:

$$g(x^{(\ell)} \mid w_0, w_1, \ldots, w_k) = w_k (x^{(\ell)})^k + \cdots + w_2 (x^{(\ell)})^2 + w_1 x^{(\ell)} + w_0$$

- Least squares estimate (assuming that $\mathbf{D}^T\mathbf{D}$ is invertible):

$$\hat{\mathbf{w}} = (\underbrace{\mathbf{D}^T\mathbf{D}}_{\mathbf{A}})^{-1} \underbrace{\mathbf{D}^T\mathbf{r}}_{\mathbf{y}}$$

where

$$\mathbf{D} = \begin{bmatrix} 1 & x^{(1)} & (x^{(1)})^2 & \cdots & (x^{(1)})^k \\ 1 & x^{(2)} & (x^{(2)})^2 & \cdots & (x^{(2)})^k \\ \vdots & & & & \\ 1 & x^{(N)} & (x^{(N)})^2 & \cdots & (x^{(N)})^k \end{bmatrix}$$

$$\mathbf{r} = (r^{(1)}, r^{(2)}, \ldots, r^{(N)})^T$$

Other Error Measures

- Squared error:
$$E(\theta \mid \mathcal{X}) = \frac{1}{2}\sum_{\ell}\left[r^{(\ell)} - g(x^{(\ell)} \mid \theta)\right]^2$$
- Relative squared error:
$$E(\theta \mid \mathcal{X}) = \frac{\sum_{\ell}\left[r^{(\ell)} - g(x^{(\ell)} \mid \theta)\right]^2}{\sum_{\ell}(r^{(\ell)} - \bar{r})^2}$$
- Absolute error:
$$E(\theta \mid \mathcal{X}) = \sum_{\ell}\left|r^{(\ell)} - g(x^{(\ell)} \mid \theta)\right|$$
- ϵ-insensitive error (for support vector machines to be covered later):
$$E(\theta \mid \mathcal{X}) = \sum_{\ell}\mathbf{1}(|r^{(\ell)} - g(x^{(\ell)} \mid \theta)| > \epsilon)(|r^{(\ell)} - g(x^{(\ell)} \mid \theta)| - \epsilon)$$

Outline

Tuning Model Complexity

- Remember that for best generalization, we should adjust the complexity of our learner model to the complexity of the data.
- In polynomial regression, the complexity parameter is the order of the fitted polynomial, and therefore we need to find a way to choose the best order that minimizes the generalization error, that is, tune the complexity of the model to best fit the complexity of the function inherent in the data.

Bias and Variance

- A sample $\mathcal{X} = \{(x^{(\ell)}, r^{(\ell)})\}_{\ell=1}^{N}$ is drawn from some unknown joint pdf $p(x, r)$.
- Using this sample $\mathcal{X}$, we construct our estimate $g(\cdot)$.
- Expected squared error of estimator $g(\cdot)$ at x over $p(x, r)$:

$$\mathbb{E}\Big[(r - g(x))^2 \mid x\Big] = \underbrace{\Big(\mathbb{E}[r \mid x] - g(x)\Big)^2}_{\text{squared error}} + \underbrace{\mathbb{E}\Big[(r - \mathbb{E}[r \mid x])^2 \mid x\Big]}_{\text{noise}}$$

Only the first term depends on the estimator $g(\cdot)$.

- The first term quantifies how much $g(x)$ deviates from the regression function, $\mathbb{E}[r \mid x]$. It depends on the estimator and the training set.
- The second term is the variance of r given x; it does not depend on $g(\cdot)$ or $\mathcal{X}$. It is the variance of noise added, σ^2. This is the part of error that can never be removed, no matter what estimator we use.

Bias and Variance (2)

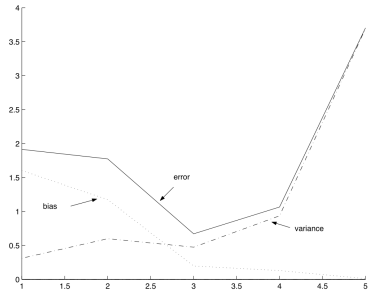
- It may be the case that for one sample, $g(x)$ may be a very good fit; and for some other sample, it may make a bad fit.
- We average over different samples $\mathcal{X}$ to quantify how well an estimator $g(\cdot)$ is, giving the expected value (average over samples $\mathcal{X}$, all of size N and drawn from the same joint density $p(x, r)$):

$$\mathbb{E}_{\mathcal{X}}\left[(\mathbb{E}[r \mid x] - g(x))^2 \mid x\right] =$$
$$\underbrace{\left(\mathbb{E}[r \mid x] - \mathbb{E}_{\mathcal{X}}[g(x)]\right)^2}_{\text{bias}^2} + \underbrace{\mathbb{E}_{\mathcal{X}}\left[(g(x) - \mathbb{E}_{\mathcal{X}}[g(x)])^2\right]}_{\text{variance}}$$

  - bias measures how much $g(x)$ is wrong disregarding the effect of varying samples
  - variance measures how much $g(x)$ fluctuate around the expected value, $\mathbb{E}_{\mathcal{X}}[g(x)]$, as the sample varies

Bias/Variance Dilemma

- As model complexity increases (e.g., order of polynomial increases):
  - bias decreases (better fit to data)
  - variance increases (fit varies more with data)
  - called the bias/variance dilemma which is true for any machine learning system
- Optimal model: best tradeoff between bias and variance.



- Model selection: search for optimal or suboptimal model.

Common Model Selection Methods

- Cross-validation: Measure generalization accuracy by testing on data unused during model training.
- Regularization: Penalize complex models by minimizing an augmented error function:

$$E' = \text{error on data} + \lambda \cdot \text{model complexity}$$

The regularization parameter λ, which can be determined by cross-validation, controls the weight of penalty.

Common Model Selection Methods (2)

- Bayesian model selection is used when there exists prior knowledge about the appropriate class of approximating functions, represented as a prior distribution over models, $p(\text{model})$.
- Posterior distribution over models given the data:

$$p(\text{model} \mid \text{data}) = \frac{p(\text{data} \mid \text{model})p(\text{model})}{p(\text{data})}$$

- From the posterior distribution, we may:
  - choose the model with the highest posterior probability, or
  - choose multiple models with high posterior probabilities, or
  - use all models weighted by their posterior probabilities.
- Taking the log of the above equation

$$\log p(\text{model} \mid \text{data}) = \log p(\text{data} \mid \text{model}) + \log p(\text{model}) - \text{const.}$$

- Regularization may be seen as a Bayesian approach in which the "prior" favors simpler models.

Common Model Selection Methods (3)

- Cross-validation is different from the other methods for model selection in that it makes no prior assumption about the model or parameters.
- If there is a large enough validation dataset, cross-validation is the best approach.
- When the data sample is small, the other models become useful.

Outline

Multivariate Data

- Multiple measurements are made on each event generating an observation vector.
- d-variate data: d inputs (a.k.a. features or attributes) – continuous or discrete.
- N i.i.d. instances (a.k.a. observations or examples) represented as a data matrix:

$$\mathbf{x} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} \\ \vdots & & & \\ x_1^{(N)} & x_2^{(N)} & \cdots & x_d^{(N)} \end{bmatrix}$$

- Classification and regression problems: predicting the value of a discrete or continuous variable, respectively.

Outline

Multivariate Parameters

- Mean vector:

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^T$$

- Covariance of x_i and x_j:

$$\sigma_{ij} = \text{Cov}(x_i, x_j) = \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)] = \mathbb{E}[x_i x_j] - \mu_i \mu_j$$

  Typically the features are correlated, or else there will not be a need for multivariate analysis.

- The covariance between two random variables measures the degree to which they are (linearly) related.
- Variance of x_i:

$$\sigma_i^2 = \mathbb{E}[(x_i - \mu_i)^2]$$

- Note that:

$$\sigma_{ij} = \sigma_{ji} \qquad \sigma_{ii} = \sigma_i^2$$

Multivariate Parameters (2)

- Covariance matrix:

$$\boldsymbol{\Sigma} \equiv \mathrm{Cov}(\mathbf{x}) = \mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T] = \mathbb{E}[\mathbf{x}\mathbf{x}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & & & \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

- Correlation between x_i and x_j:

$$\rho_{ij} \equiv \mathrm{Corr}(x_i, x_j) = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

  The correlation (a.k.a. Pearson correlation coefficient) between x_i and x_j is in $[-1, +1]$, making it easier to interpret than the covariance.
  - $\rho_{ij} \neq 0$: two variables x_i and x_j are related in a linear way
- Dependence vs. correlation:

$$x_i \text{ and } x_j \text{ are independent } \overset{\Rightarrow}{\not\Leftarrow} \sigma_{ij} = \rho_{ij} = 0$$

Parameter Estimation

- Sample mean (ML estimator of $\boldsymbol{\mu}$):

$$\mathbf{m} = \frac{1}{N}\sum_{\ell=1}^{N}\mathbf{x}^{(\ell)}$$

- Sample covariance matrix (ML estimator of $\boldsymbol{\Sigma}$):

$$\mathbf{S} = [s_{ij}]_{i,j=1}^{d} = \frac{1}{N}\sum_{\ell=1}^{N}(\mathbf{x}^{(\ell)} - \mathbf{m})(\mathbf{x}^{(\ell)} - \mathbf{m})^T$$

  where $s_{ii} = s_i^2$

- Sample correlation matrix:

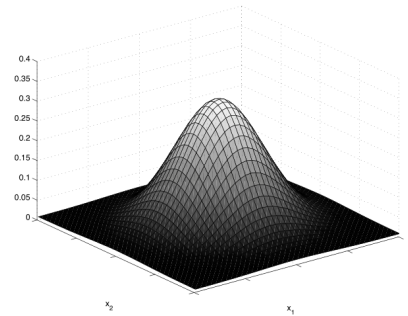
$$\mathbf{R} = [r_{ij}]_{i,j=1}^{d} \quad \text{where} \quad r_{ij} = \frac{s_{ij}}{s_i s_j}$$

Estimation of Missing Values

- What to do if the values of certain variables in some instances are missing?
- Discarding the instances: not a good idea if the sample is small and since the non-missing entries do contain information.
- Imputation: filling in the missing entries
  - Mean imputation: using the most likely value (e.g., mean or mode)
  - Imputation by regression: predicting the missing values based on the regression approach
  - Matrix factorization: using low-rank matrices as factors for matrix completion.
  - sometimes a certain missing attribute value may be important; it can be represented as a separate value to indicate that the value is missing and is used as such.

Outline

Multivariate Normal Distribution



$$\mathbf{x} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\Big[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big]$$

Multivariate Normal Distribution (2)

- Multivariate generalization of univariate normal distribution.
- Multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $d \times 1$ mean vector $\boldsymbol{\mu}$ and $d \times d$ covariance matrix $\boldsymbol{\Sigma}$.
- Probability density function:

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\Big[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\Big]$$

- Log likelihood:

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{x}) = -\frac{Nd}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{\ell=1}^{N}(\mathbf{x}^{(\ell)} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}^{\ell} - \boldsymbol{\mu})$$

- ML estimates:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{\ell=1}^{N} \boldsymbol{x}^{(\ell)} \qquad \hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{\ell=1}^{N}(\mathbf{x}^{(\ell)} - \hat{\boldsymbol{\mu}})(\mathbf{x}^{(\ell)} - \hat{\boldsymbol{\mu}})^T$$

Multivariate Normal Distribution (3)

- Mahalanobis distance measures the distance from $\mathbf{x}$ to $\boldsymbol{\mu}$ in terms of $\boldsymbol{\Sigma}$ (normalized for differences in variance and covariance):

$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

- $(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2$ is the *d*-dimensional hyperellipsoid centered at $\boldsymbol{\mu}$. Its shape and orientation are defined by $\boldsymbol{\Sigma}$.
- Euclidean distance is a special case of Mahalanobis distance when $\boldsymbol{\Sigma} = s^2\mathbf{I}$; the hyperellipsoid degenerates into a hypersphere.

Bivariate Normal Distribution

- Multivariate normal distribution with $d = 2$.
- Covariance matrix:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{bmatrix}$$

- Joint density:

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\Big[-\frac{1}{2(1-\rho^2)}(z_1^2 - 2\rho z_1 z_2 + z_2^2)\Big]$$

where

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \ (z\text{-normalization})$$

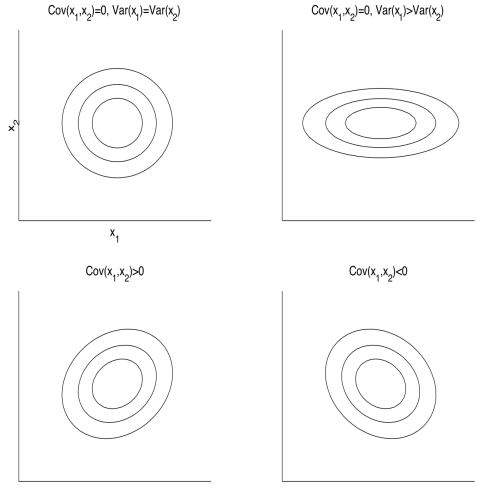
Bivariate Normal Distribution (2)

- for $|\rho| < 1$, the equation of an ellipse

$$z_1^2 - 2\rho z_1 z_2 + z_2^2 = c^2$$

  - if $\rho > 0$, the major axis of the ellipse has a positive slope
  - if $\rho < 0$, the major axis of the ellipse has a negative slope
  - If $\rho = 0$, the two variables are independent, the cross-term disappears, and we get a product of two univariate densities.
- If $\rho = \pm 1$, the two variables are linearly related, the observations are effectively one-dimensional, and one of the two variables can be disposed of.

Isoprobability Contour Plot of Bivariate Normal



Independent Inputs

- If x_i are independent, the off-diagonal entries σ_{ij}, $i \neq j$ of $\mathbf{\Sigma}$ are 0. The joint density becomes:

$$p(\mathbf{x}) = \prod_{i=1}^{d} p_i(x_i) = \frac{1}{(2\pi)^{d/2} \prod_{i=1}^{d} \sigma_i} \exp\left[-\frac{1}{2} \sum_{i=1}^{d} \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right]$$

  Mahalanobis distance reduces to weighted Euclidean distance (with weightings $1/\sigma_i$).
- It further reduces to Euclidean distance if all variances σ_i^2 are equal.

Outline

Parametric Classification

- In Bayes' decision rule for classification, the discriminant function for of class C_i is

$$p(\mathbf{x} \mid C_i)P(C_i) \quad \text{or} \quad \log\left[p(\mathbf{x} \mid C_i)P(C_i)\right]$$

- Class-conditional densities $p(\mathbf{x} \mid C_i) \sim \mathcal{N}_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$:

$$p(\mathbf{x} \mid C_i) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\Big[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\Big]$$

- Discriminant functions:

$$\begin{aligned} g_i(\mathbf{x}) &= \log p(\mathbf{x} \mid C_i) + \log P(C_i) \\ &= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \log P(C_i) \end{aligned}$$

Estimation of Parameters

- Given a training sample for $K \geq 2$ classes, $\mathcal{X} = \{\mathbf{x}^{(\ell)}, \mathbf{r}^{(\ell)}\}$, where $r_i^{(\ell)} = 1$ if $\mathbf{x}^{(\ell)} \in C_i$ and 0 otherwise, parameters can be estimated separately for each class.
- Parameter estimates:

$$\hat{P}(C_i) = \frac{1}{N} \sum_{\ell} r_i^{(\ell)}$$

$$\mathbf{m}_i = \frac{\sum_{\ell} r_i^{(\ell)} \mathbf{x}^{(\ell)}}{\sum_{\ell} r_i^{(\ell)}}$$

$$\mathbf{S}_i = \frac{\sum_{\ell} r_i^{(\ell)} (\mathbf{x}^{(\ell)} - \mathbf{m}_i)(\mathbf{x}^{(\ell)} - \mathbf{m}_i)^T}{\sum_{\ell} r_i^{(\ell)}}$$

Quadratic Discriminant Functions

- The parameter estimates are then plugged into the discriminant functions:

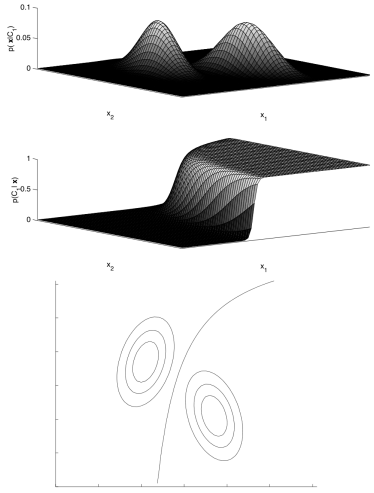
$$\begin{aligned} g_i(\mathbf{x}) &= -\frac{1}{2}\log|\mathbf{S}_i| - \frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{S}_i^{-1}(\mathbf{x}-\mathbf{m}_i) + \log\hat{P}(C_i) \\ &= -\frac{1}{2}\log|\mathbf{S}_i| - \frac{1}{2}(\mathbf{x}^T\mathbf{S}_i^{-1}\mathbf{x} - 2\mathbf{x}^T\mathbf{S}_i^{-1}\mathbf{m}_i + \mathbf{m}_i^T\mathbf{S}_i^{-1}\mathbf{m}_i) + \log\hat{P}(C_i) \\ &= \mathbf{x}^T\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^T\mathbf{x} + w_{i0} \end{aligned}$$

where

$$\begin{aligned} \mathbf{W}_i &= -\frac{1}{2}\mathbf{S}_i^{-1} \\ \mathbf{w}_i &= \mathbf{S}_i^{-1}\mathbf{m}_i \\ w_{i0} &= -\frac{1}{2}\mathbf{m}_i^T\mathbf{S}_i^{-1}\mathbf{m}_i - \frac{1}{2}\log|\mathbf{S}_i| + \log\hat{P}(C_i) \end{aligned}$$

- The discriminant functions are quadratic.

Quadratic Discriminant Functions (2)



Quadratic Discriminant Functions (3)

- The number of parameters to be estimated are Kd for the means and $Kd(d+1)/2$ for the covariance matrices.
- When d is large and samples are small, the estimation is not reliable.
- For the estimates to be reliable on small samples, one may want to decrease dimensionality, d, by redesigning the feature extractor and select a subset of the features or somehow combine existing features.
- Another possibility is to pool the data and estimate a common covariance matrix for all classes.

Common Covariance Matrix S

- Shared common sample covariance matrix:

$$\mathbf{S} = \sum_i \hat{P}(C_i)\mathbf{S}_i$$

- Discriminant functions are linear:

$$\begin{aligned} g_i(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{S}^{-1}(\mathbf{x}-\mathbf{m}_i) + \log \hat{P}(C_i) + \text{const.} \\ &= \mathbf{w}_i^T\mathbf{x} + w_{i0} \quad \text{(ignoring terms that are the same for all classes)} \end{aligned}$$

where

$$\mathbf{w}_i = \mathbf{S}^{-1}\mathbf{m}_i$$

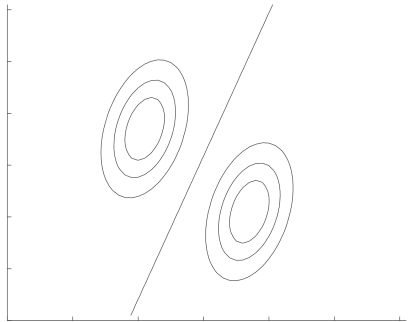
$$w_{i0} = -\frac{1}{2}\mathbf{m}_i^T\mathbf{S}^{-1}\mathbf{m}_i + \log \hat{P}(C_i)$$

- The number of parameters is Kd for the means and $d(d+1)/2$ for the shared covariance matrix.

Common Covariance Matrix S (2)

- If the priors are equal, the optimal decision rule is to assign input to the class whose mean's Mahalanobis distance to the input is the smallest.
- Unequal priors shift the boundary toward the less likely class.

Common Covariance Matrix S (3)



Decision regions of such a linear classifier are convex.

Diagonal S

- Naive Bayes' classifier: if the variables are independent, $\mathbf{\Sigma}$ becomes a diagonal matrix.
- Class-conditional densities:

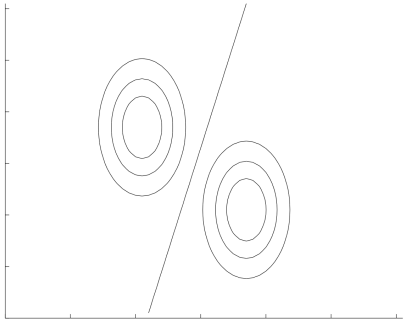
$$p(\mathbf{x} \mid C_i) = \prod_j p(x_j \mid C_i)$$

  where $p(x_j \mid C_i)$ are univariate Gaussian distributions.
- Discriminant functions:

$$g_i(\mathbf{x}) = -\frac{1}{2}\sum_{j=1}^{d}\left(\frac{x_j - m_{ij}}{s_j}\right)^2 + \log \hat{P}(C_i)$$

- Classification based on weighted Euclidean distance.
- The number of parameters is Kd for the means and d for the variances.

Diagonal S (2)



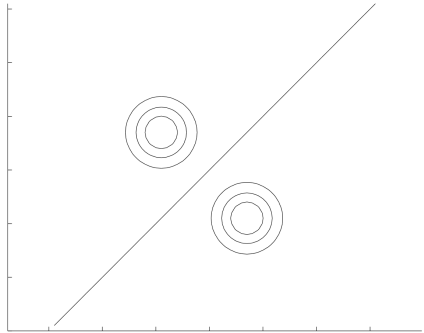
Diagonal S with Equal Variances

- If we assume further that all variances are equal, i.e., $\mathbf{\Sigma} = s^2\mathbf{I}$, weighted Euclidean distance reduces to Euclidean distance.
- Discriminant functions:

$$\begin{aligned} g_i(\mathbf{x}) &= -\frac{1}{2s^2}\|\mathbf{x} - \mathbf{m}_i\|^2 + \log \hat{P}(C_i) \\ &= -\frac{1}{2s^2}\sum_{j=1}^{d}(x_j - m_{ij})^2 + \log \hat{P}(C_i) \end{aligned}$$

- The number of parameters in this case is Kd for the means and 1 for s^2.
- If the priors are equal, we have $g_i(\mathbf{x}) = -\|\mathbf{x} - \mathbf{m}_i\|^2$
  - nearest mean classifier: it assigns the input to the class of the nearest mean
  - template matching procedure: each mean acts as a prototype or template for the class.

Diagonal S with Equal Variances (2)



Tuning Model Complexity

Assumption	Covariance matrix	No. of parameters
Shared, hyperspherical	$\mathbf{S}_i = \mathbf{S} = s^2\mathbf{I}$	1
Shared, axis-aligned	$\mathbf{S}_i = \mathbf{S}$, with $s_{ij} = 0$	d
Shared, hyperellipsoidal	$\mathbf{S}_i = \mathbf{S}$	$d(d+1)/2$
Different, hyperellipsoidal	$\mathbf{S}_i$	$Kd(d+1)/2$

- Complexity increases (i.e., less restricted $\mathbf{S}$)
  $\Rightarrow$ bias decreases and variance increases
- Regularization: uses strong bias to control model complexity.

Outline

Discrete Features: Binary

- Binary (Bernoulli) variables x_j:

$$p_{ij} \equiv p(x_j = 1 \mid C_i)$$

- If x_j are independent given C_i (naive Bayes'):

$$p(\mathbf{x} \mid C_i) = \prod_{j=1}^{d} p_{ij}^{x_j} (1 - p_{ij})^{1 - x_j}$$

giving linear discriminant functions:

$$\begin{aligned} g_i(\mathbf{x}) &= \log p(\mathbf{x} \mid C_i) + \log P(C_i) \\ &= \sum_j \Big[x_j \log p_{ij} + (1 - x_j) \log(1 - p_{ij}) \Big] + \log P(C_i) \end{aligned}$$

- Estimated parameters:

$$\hat{p}_{ij} = \sum_{\ell} x_j^{(\ell)} r_i^{(\ell)} \Big/ \sum_{\ell} r_i^{(\ell)}$$

Discrete Features: Multinomial

- Multinomial variables $x_j \in \{v_1, \ldots, v_{n_j}\}$
- Indicator variables:

$$z_{jk} = \begin{cases} 1 & \text{if } x_j = v_k \\ 0 & \text{otherwise} \end{cases}$$

- Define

$$p_{ijk} \equiv p(z_{jk} = 1 \mid C_i) = p(x_j = v_k \mid C_i)$$

- If x_j are independent:

$$p(\mathbf{x} \mid C_i) = \prod_{j=1}^{d} \prod_{k=1}^{n_j} p_{ijk}^{z_{jk}}$$

$$g_i(\mathbf{x}) = \sum_j \sum_k z_{jk} \log p_{ijk} + \log P(C_i)$$

- The maximum likelihood estimator

$$\hat{p}_{ijk} = \frac{\sum_\ell z_{jk}^{(\ell)} r_i^{(\ell)}}{\sum_\ell r_i^{(\ell)}}$$

Outline

Multivariate Regression

- Multivariate linear regression:

$$\begin{aligned} r &= g(\mathbf{x} \mid w_0, w_1, \ldots, w_d) + \epsilon \\ &= w_0 + w_1 x_1 + \cdots + w_d x_d + \epsilon \end{aligned}$$

  In statistical literature, this is called multiple regression; statisticians use the term multivariate when there are multiple outputs.

- Error function:

$$E(w_0, w_1, \ldots, w_d \mid \mathcal{X}) = \frac{1}{2} \sum_{\ell} \left(r^{(\ell)} - w_0 - w_1 x_1^{(\ell)} - \cdots - w_d x_d^{(\ell)} \right)^2$$

- Maximizing the Gaussian likelihood is equivalent to minimizing the sum of squared errors.

Normal Equations

- Taking the derivative with respect to the parameters, we get the normal equations for multivariate linear regression:

$$\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{r}$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} \\ \vdots & & & & \\ 1 & x_1^{(N)} & x_2^{(N)} & \cdots & x_d^{(N)} \end{bmatrix}$$

$$\mathbf{w} = (w_0, w_1, \ldots, w_d)^T$$

$$\mathbf{r} = (r^{(1)}, r^{(2)}, \ldots, r^{(N)})^T$$

- Estimated parameters (assuming that $\mathbf{X}^T\mathbf{X}$ is invertible):

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{r}$$

Multivariate Polynomial Regression

- Define new higher-order variables, e.g.

$$z_1 = x_1,\ z_2 = x_2\ z_3 = (x_1)^2,\ z_4(x_2)^2,\ z_5 = x_1 x_2$$

- Apply multivariate linear regression in the new $\mathbf{z}$ space.
- Actually using higher-order terms of inputs as additional inputs is only one possibility; we can define any nonlinear function of the original inputs using basis functions, like $z = \sin(x)$.
- This idea of generalizing the linear model is frequently used in later course.