Bayesian Decision Theory and Parameter Estimation

Hong Chang

Institute of Computing Technology,
Chinese Academy of Sciences

Pattern Recognition and Machine Learning (Fall 2021)

Outline I

Coin Tossing Example

- Outcome of tossing a coin $\in \{\text{head,tail}\}$
- Random variable X:

$$X = \begin{cases} 1 & \text{if outcome is head} \\ 0 & \text{if outcome is tail} \end{cases}$$

- X is Bernoulli-distributed:

$$P(X) = p_0^X (1 - p_0)^{1-X}$$

where the parameter p_0 is the probability that the outcome is head, i.e., $p_0 = P(X = 1)$.

Estimation and Prediction

- Estimation of parameter p_0 from sample $\mathcal{X} = \{x^{(i)}\}_{i=1}^N$:

$$\hat{p}_0 = \frac{\sharp heads}{\sharp tosses} = \frac{\sum_{i=1}^N x^{(i)}}{N}$$

- Prediction of outcome of next toss:

$$\text{Predicted outcome} = \begin{cases} \text{head} & \text{if } p_0 > 1/2 \\ \text{tail} & \text{otherwise} \end{cases}$$

by choosing the more probable outcome, which minimizes the probability of error (=1-probability of our choice for the predicted outcome).

Classification as Bayesian Decision

- Credit scoring example:
  - Inputs: income and savings, or $\mathbf{x} = (x_1, x_2)^T$
  - Output: risk $\in \{\text{low,high}\}$, or $C \in \{0, 1\}$
- Prediction:

$$\text{Choose} = \begin{cases} C = 1 & \text{if } P(C = 1|\mathbf{x}) > 0.5 \\ C = 0 & \text{otherwise} \end{cases}$$

  or equivalently

$$\text{Choose} = \begin{cases} C = 1 & \text{if } P(C = 1|\mathbf{x}) > P(C = 0|\mathbf{x}) \\ C = 0 & \text{otherwise} \end{cases}$$

- Probability of error:

$$1 - \max(P(C = 1|\mathbf{x}), P(C = 0|\mathbf{x}))$$

Bayes' Rule

- Bayes' rule:

$$\text{Posterior } P(C|\mathbf{x}) = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} = \frac{p(\mathbf{x}|C)P(C)}{p(\mathbf{x})}$$

- Some useful properties:
  - $P(C = 1) + P(C = 0) = 1$
  - $p(\mathbf{x}) = p(\mathbf{x}|C = 1)P(C = 1) + p(\mathbf{x}|C = 0)P(C = 0)$
  - $P(C = 0|\mathbf{x}) + P(C = 1|\mathbf{x}) = 1$

Bayes' Rule for Multiple Classes

- Bayes' rule for general case (K mutually exclusive and exhaustive classes):

$$
\begin{aligned}
P(C_i|\mathbf{x}) &= \frac{p(\mathbf{x}|C_i)P(C_i)}{p(\mathbf{x})} \\
&= \frac{p(\mathbf{x}|C_i)P(C_i)}{\sum_{k=1}^{K} p(\mathbf{x}|C_k)P(C_k)}
\end{aligned}
$$

- Optimal decision rule for Bayes' classifier:

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

Losses and Risks

- Different decisions or actions may not be equally good or costly.
- Action α_i: decision to assign the input $\mathbf{x}$ to class C_i
- Loss λ_{ik}: loss incurred for taking action α_i when the actual state if C_k
- Expected risk for taking action α_i:

$$R(\alpha_i|\mathbf{x}) = \sum_{k=1}^{K} \lambda_{ik} P(C_k|\mathbf{x})$$

- Optimal decision rule with minimum expected risk:

$$\text{Choose } \alpha_i \text{ if } R(\alpha_i|\mathbf{x}) = \min_k R(\alpha_k|\mathbf{x})$$

0-1 Loss

- All correct decisions have no loss and all errors have unit cost:

$$\lambda_{ik} = \begin{cases} 0 & \text{if } i = k \\ 1 & \text{if } i \neq k \end{cases}$$

- Expected risk:

$$\begin{aligned} R(\alpha_i|\mathbf{x}) &= \sum_{k=1}^{K} \lambda_{ik} P(C_k|\mathbf{x}) \\ &= \sum_{k \neq i} P(C_k|\mathbf{x}) = 1 - P(C_i|\mathbf{x}) \end{aligned}$$

- Optimal decision rule with minimum expected risk (or, equivalently, highest posterior probability):

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

Discriminant Functions

- One way of performing classification is through a set of discriminant functions.
- Classification rule:

$$\text{Choose } C_i \text{ if } g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})$$

- Different ways of defining the discriminant functions:
  - $g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$
  - $g_i(\mathbf{x}) = P(C_i|\mathbf{x})$
  - $g_i(\mathbf{x}) = p(\mathbf{x}|C_i)P(C_i)$
- For the two-class case, we may define a single discriminant function:

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$$

  with the following classification rule:

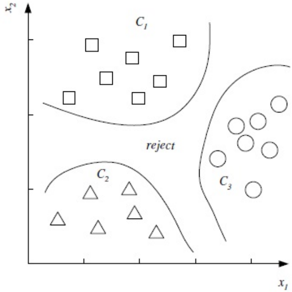
$$\text{Choose} \begin{cases} C_1 & \text{if } g(\mathbf{x}) > 0 \\ C_2 & \text{otherwise} \end{cases}$$

Decision Regions

- The feature space is divided into K decision regions $\mathcal{R}_1, \ldots, \mathcal{R}_K$, where

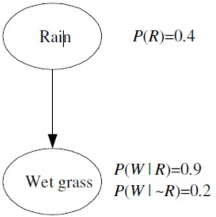
$$\mathcal{R}_i = \{\mathbf{x} | g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})\}$$

- The decision regions are separated by decision boundaries where ties occur among the largest discriminant functions.

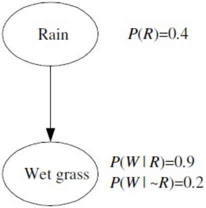


Bayesian Networks

- A.k.a. belief networks, probabilistic networks, or more generally graphical models.
- Node (or vertex): random variable
- Edge (or arc or link): direct influence between variables
- Structure: directed acyclic graph (DAG) formed by nodes and edges
- Parameters: probabilities and conditional probabilities



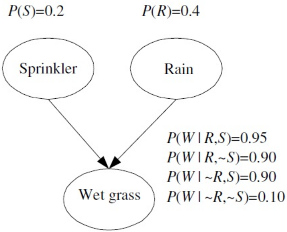
Causal Graph and Diagnostic Inference



- Causal graph: rain is the cause of wet grass.
- Diagnostic inference: knowing that the grass is wet, what is the probability that rain is the cause?
- Bayes' rule:

$$
\begin{aligned}
P(R|W) &= \frac{P(W|R)P(R)}{P(W)} \\
&= \frac{0.9 \times 0.4}{0.9 \times 0.4 + 0.2 \times 0.6} = 0.75 > P(R) = 0.4
\end{aligned}
$$

Two Causes: Causal Inference



- Causal or predictive inference: if the sprinkler is on, what is the probability that the grass is wet?

$$
\begin{aligned}
P(W|S) &= P(W|R,S)P(R|S) + P(W|\sim R,S)P(\sim R|S) \\
&= P(W|R,S)P(R) + P(W|\sim R,S)P(\sim R) \\
&= 0.95 \times 0.4 + 0.9 \times 0.6 \\
&= 0.92
\end{aligned}
$$

Two Causes: Diagnostic Inference

- Diagnostic inference: if the grass is wet, what is the probability that the sprinkler is on?

$$P(S|W) = \frac{P(W|S)P(S)}{P(W)}$$

where

$$\begin{aligned} P(W) &= P(W|R,S)P(R,S) + P(W|\sim R,S)P(\sim R,S) + \\ &\quad P(W|R,\sim S)P(R,\sim S) + P(W|\sim R,\sim S)P(\sim R,\sim S) \\ &= P(W|R,S)P(R)P(S) + P(W|\sim R,S)P(\sim R)P(S) + \\ &\quad P(W|R,\sim S)P(R)P(\sim S) + P(W|\sim R,\sim S)P(\sim R)P(\sim S) \\ &= 0.52 \end{aligned}$$

so

$$P(S|W) = \frac{0.92 \times 0.2}{0.52} = 0.35 > P(S) = 0.2$$

Two Causes: Diagnostic Inference

- Diagnostic inference: given rain and wet grass, what is the probability that the sprinkler is on?

$$P(S|R, W) = ?$$

- Note that

$$P(S|R, W) = \frac{P(W|R, S)P(S|R)}{P(W|R)} = \frac{P(W|R, S)P(S)}{P(W|R)}$$

Two Causes: Explaining Away

- Bayes' rule:

$$P(S|R,W) = \frac{P(W|R,S)P(S|R)}{P(W|R)} = \frac{P(W|R,S)P(S)}{P(W|R)} = 0.21$$

- Explaining away:

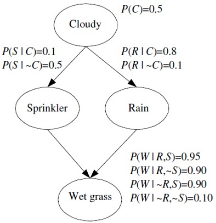
$$0.21 = P(S|R,W) < P(S|W) = 0.35$$

Knowing that it has rained decreases that probability that the sprinkler is on.

- Knowing that the grass is wet, rain and sprinkler become dependent:

$$P(S|R,W) \neq P(S|W)$$

Dependent Causes

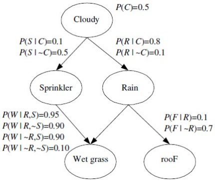


- Causal inference:

$$\begin{aligned} P(W|C) &= P(W|R,S,C)P(R,S|C) + P(W|\sim R,S,C)P(\sim R,S|C) + \\ &\quad P(W|R,\sim S,C)P(R,\sim S|C) + P(W|\sim R,\sim S,C)P(\sim R,\sim S|C) \\ &= P(W|R,S)P(R|C)P(S|C) + P(W|\sim R,S)P(\sim R|C)P(S|C) + \\ &\quad P(W|R,\sim S)P(R|C)P(\sim S|C) + P(W|\sim R,\sim S)P(\sim R|C)P(\sim S|C) \end{aligned}$$

- Independence: W and C are independent given R and S; R and S are independent given C.

Local Structures



- The network represents conditional independence statements.
- The joint distribution can be broken down into local structures:

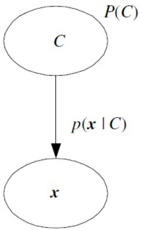
$$P(C, S, R, W, F) = P(C)P(S|C)P(R|C)P(W|S, R)P(F|R)$$

- In general,

$$P(X_1, \ldots, X_d) = \prod_{i=1}^{d} P(X_i|parents(X_i))$$

where X_i is either continuous or discrete with ≥ 2 possible values.

Bayesian Network for Classification

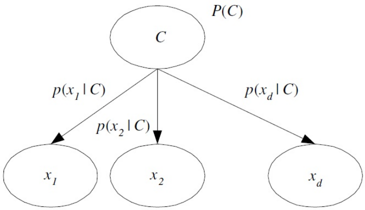


- Bayes' rule inverts the edge:

$$P(C|\mathbf{x}) = \frac{P(\mathbf{x}|C)P(C)}{P(\mathbf{x})}$$

- Classification as diagnostic inference.

Naive Bayes' Classifier



- Given C, the input variables x_j are independent:

$$p(\mathbf{x}|C) = \prod_{j=1}^{d} p(x_j|C)$$

- The Naive Bayes' classifier ignores possible dependencies among the input variables and reduces a multivariate problem to a group of univariate problems.

Maximum Likelihood Estimation

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

$$L(\boldsymbol{\theta}|\mathcal{X}) \doteq p(\mathcal{X}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})$$

- Log likelihood (mainly for computational simplification):

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{X}) \doteq \log L(\boldsymbol{\theta}|\mathcal{X}) = \sum_{i=1}^{N} \log p(\mathbf{x}^{(i)}|\boldsymbol{\theta})$$

- Maximum likelihood estimate:

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

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 distribution on over x (with parameter $\theta = p_0$):

$$P(x|p_0) = p_0^x (1 - p_0)^{1-x}$$

- Log likelihood:

$$\mathcal{L}(p_0|\mathcal{X}) = \sum_{i=1}^{N} x^{(i)} \log p_0 + (1 - x^{(i)}) \log(1 - p_0)$$

- ML estimate:

$$\hat{p}_0 = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}$$

Example: Multinomial

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

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

- Probability distribution function (with parameters $\boldsymbol{\theta} = (p_1, \ldots, p_K)^T$):

$$P(\mathbf{x}|\boldsymbol{\theta}) = P(x_1, \ldots, x_K | p_1, \ldots, p_K) = \prod_{k=1}^{K} p_k^{x_k}$$

with constraint $\sum_{k=1}^{K} p_k = 1$.

Example: Multinomial (2)

- Log likelihood:

$$\mathcal{L}(p_1, \ldots, p_K | \mathcal{X}) = \sum_{i=1}^{N} \sum_{k=1}^{K} x_k^{(i)} \log p_k$$

- ML estimates:

$$\hat{p}_k = \frac{1}{N} \sum_{i=1}^{N} x_k^{(i)}$$

So $\hat{\theta} = (\hat{p}_1, \ldots, \hat{p}_K)^T$ is also the sample mean.

Example: Normal

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

$$p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

- Log likelihood:

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

- ML estimates:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}, \quad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x^{(i)} - \hat{\mu})^2$$

Example: Multivariate Normal

- Multivariate generalization of univariate normal distribution.
- Multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^2)$ with mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and variance $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$.
- Probability density function:

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

- Log likelihood:

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

- ML estimates:

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

Bayesian Estimation

- Unlike MLE which treats θ as a fixed (but unknown) point, the Bayesian approach treats it as a random variable with prior density $p(\theta)$ modeling the prior uncertainty about θ.
- Posterior density θ (uncertainty about θ after observing the sample):

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

- Full Bayesian approach:
  - Estimation of density at x:

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

  - Prediction (e.g., regression) in the form $y = f(x|\theta)$:

$$y = \int f(x|\theta)p(\theta|\mathcal{X})d\theta$$

Computational Considerations

- Evaluating the integrals may be difficult, so the full Bayesian approach may be replaced by some other methods.
- Maximum a posteriori (MAP) estimation:

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

  $p(x|\mathcal{X}) \approx p(x|\theta_{MAP})$ and $y \approx y_{MAP} = f(x|\theta_{MAP})$.
- Maximum likelihood (ML) estimation - MAP with flat prior:

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

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

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

Example: Bayesian Estimation for Gaussian

- Bayesian estimation for μ, with known μ_0,σ and σ_0:

$$\begin{aligned} x^{(i)} &\sim \mathcal{N}(\mu, \sigma^2) \\ \mu &\sim \mathcal{N}(\mu_0, \sigma_0^2) \end{aligned}$$

- The likelihood (given training set $\mathcal{X}$):

$$p(\mathcal{X}|\mu) = \prod_{i=1}^{N} p(x^{(i)}|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x^{(i)} - \mu)^2\right)$$

- The prior:

$$p(\mu) = \frac{1}{(2\pi\sigma_0^2)^{1/2}} \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right)$$

- The posterior distribution:

$$p(\mu|\mathcal{X}) \propto p(\mathcal{X}|\mu)p(\mu)$$

Example: Bayesian Estimation for Gaussian (2)

- MLE:

$$\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)} \quad \text{(sample mean)}$$

- The posterior distribution is given:

$$p(\mu|\mathcal{X}) = \mathcal{N}(\mu_N, \sigma_N^2)$$

where

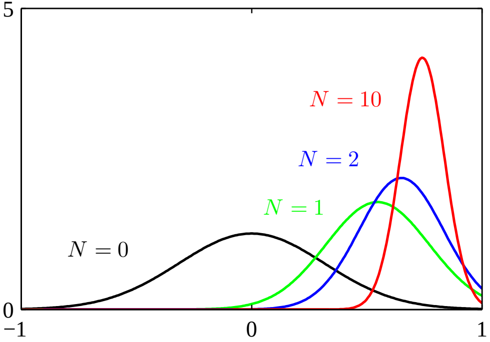
$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML} \quad (1)$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \quad (2)$$

  - The mean of posterior distribution μ_N is a weighted average of sample mean μ_{ML} and prior mean μ_0.

Illustration

- The data points are generated from a Gaussian of mean 0.8 and variance 1.
- The prior is chosen to have mean 0.
- The variance is set to the true value.



Parametric Approach

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

Parametric Approach to Classification

- Recall Bayes' rule for classification:

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

- $p(\mathbf{x}|C_i)$ and $P(C_i)$ need to be estimated from the sample $\mathcal{X}$.

Classification with discriminant Functions

- Gaussian density for each class:

$$p(x|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp(-\frac{(x-\mu_k)^2}{2\sigma_k^2})$$

- Discriminant functions:

$$\begin{aligned} g_k(x) &= \log(p(x|C_k)P(C_k)) \\ &= \log p(x|C_k) + \log P(C_k) \\ &= -\frac{1}{2}\log 2\pi - \log \sigma_k - \frac{(x-\mu_k)^2}{2\sigma_k^2} + \log P(C_k) \end{aligned}$$

Discriminant Functions Based on ML Estimates

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

$$x^{(i)} \in \mathbb{R}, \quad y_k^{(i)} = \begin{cases} 1 & \text{if } x^{(i)} \text{ belongs to } C_k \\ 0 & \text{if } x^{(i)} \text{ belongs to } C_j, j \neq k \end{cases}$$

- ML estimates:

$$\hat{P}(C_k) = \frac{1}{N}\sum_{i=1}^{N} y_k^{(i)}$$

$$m_k = \frac{\sum_{i=1}^{N} x^{(i)} y_k^{(i)}}{\sum_{i=1}^{N} y_k^{(i)}}, \quad s_k^2 = \frac{\sum_{i=1}^{N} (x^{(i)} - m_k)^2 y_k^{(i)}}{\sum_{i=1}^{N} y_k^{(i)}}$$

- Discriminant function (dropping constant term):

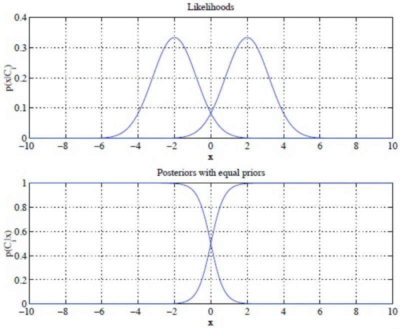
$$g_k(x) = -\log s_k - \frac{(x - m_k)^2}{2s_k^2} + \log \hat{P}(C_k)$$

Special Case: Equal Priors and Variances

- Simplified discriminant functions: $g_k(x) = -(x - m_k)^2$
- Classification rule (nearest mean classifier):

$$\text{Choose } C_k \text{ if } |x - m_k| = \min_j |x - m_j|$$

- Likelihood function and posterior densities:

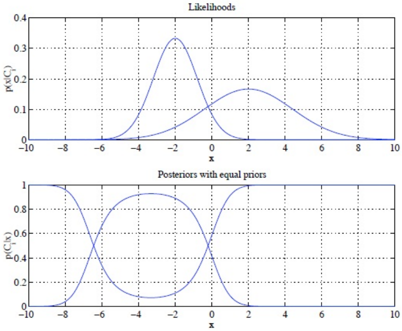


Special Case: Equal Priors but Different Variances

- Simplified discriminant functions:

$$g_k(x) = -\log s_k - \frac{(x - m_k)^2}{2s_k^2}$$

- Likelihood functions and posterior densities:



Additive Parametric Model

- Parametric modeling:

$$y = f_{\mathbf{w}}(x) + \epsilon$$

  - $f_{\mathbf{w}}(x)$ is the estimate regression function with parameters $\mathbf{w}$
  - $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is random noise independent of the input

$$p(\epsilon) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{\epsilon^2}{2\sigma^2})$$

- Conditional probability output given input $p(y|x) \sim \mathcal{N}(f_{\mathbf{w}}(x), \sigma^2)$, i.e.,

$$p(y^{(i)}|x^{(i)}; \mathbf{w}) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y^{(i)} - f_{\mathbf{w}}(x^{(i)}))^2}{2\sigma^2})$$

Maximum Likelihood Estimation

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

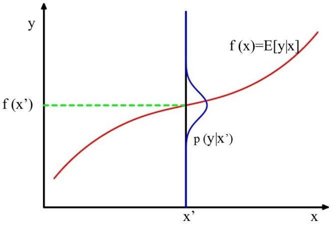
$$
\begin{aligned}
\mathcal{L}(\mathbf{w}|\mathcal{X}) &= \sum_{i=1}^{N} \log p(y^{(i)}|x^{(i)}; \mathbf{w}) \\
&= \log \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - f_{\mathbf{w}}(x^{(i)}))^2}{2\sigma^2}\right) \\
&= N \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^{N} (y^{(i)} - f_{\mathbf{w}}(x^{(i)}))^2
\end{aligned}
$$

Maximum Likelihood Estimation (2)

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

$$E(\mathbf{w}|\mathcal{X}) = \frac{1}{2}\sum_{i=1}^{N}(y^{(i)} - f_{\mathbf{w}}(x^{(i)}))^2$$

  - So the ML estimate of $\mathbf{w}$ is also called the least square estimate.
- $f_{\mathbf{w}}(x)$ is given by the mean of the conditional distribution $p(y|x)$



Example: Linear Regression

- Linear regression function:

$$f(x^{(i)}|w_0, w_1) = w_1 x^{(i)} + w_0$$

- Error function:

$$E(w_0, w_1|\mathcal{X}) = \frac{1}{2}\sum_{i=1}^{N}(y^{(i)} - w_1 x^{(i)} - w_0)^2$$

- Setting the derivatives of $E(w_0, w_1|\mathcal{X})$ w.r.t. w_0, w_1 to 0 gives:

$$\begin{aligned}\sum_{i=1}^{N} y^{(i)} &= N w_0 + w_1 \sum_{i=1}^{N} x^{(i)} \\ \sum_{i=1}^{N} y^{(i)} x^{(i)} &= w_0 \sum_{i=1}^{N} x^{(i)} + w_1 \sum_{i=1}^{N} (x^{(i)})^2\end{aligned}$$

Example: Linear Regression (2)

- Linear system in matrix form:

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

where

$$\begin{aligned}
\mathbf{A} &= \begin{bmatrix} N & \sum_i x^{(i)} \\ \sum_i x^{(i)} & \sum_i (x^{(i)})^2 \end{bmatrix} \\
\mathbf{w} &= \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \\
\mathbf{z} &= \begin{pmatrix} \sum_i y^{(i)} \\ \sum_i y^{(i)} x^{(i)} \end{pmatrix}
\end{aligned}$$

- Least squares estimate:

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

Example: Polynomial Regression

- Polynomial regression function of order m:

$$f(x^{(i)}|w_0, \ldots, w_m) = w_m(x^{(i)})^m + \ldots + w_1 x^{(i)} + w_0$$

- Least squares estimate:

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

where

$$\mathbf{D} = \begin{pmatrix} 1 & x^{(1)} & (x^{(1)})^2 & \ldots & (x^{(1)})^m \\ 1 & x^{(2)} & (x^{(2)})^2 & \ldots & (x^{(2)})^m \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x^{(N)} & (x^{(N)})^2 & \ldots & (x^{(N)})^m \end{pmatrix}$$

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

Generalized Linear Regression

- Linear regression with of nonlinear basis functions:

$$f_{\mathbf{w}}(x) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(x) = \mathbf{w}^T \phi(x)$$

  - $\phi_0(x) = 1$, $\phi(x) = [\phi_0(x), \phi_1(x), \ldots, \phi_{M-1}(x)]^T$
- Define $N \times M$ matrix $\mathbf{\Phi}$:

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(x^{(1)}) & \phi_1(x^{(1)}) & \cdots & \phi_{M-1}(x^{(1)}) \\ \phi_0(x^{(2)}) & \phi_1(x^{(2)}) & \cdots & \phi_{M-1}(x^{(2)}) \\ \vdots & \vdots & \ddots & \cdots \\ \phi_0(x^{(N)}) & \phi_1(x^{(N)}) & \cdots & \phi_{M-1}(x^{(N)}) \end{pmatrix}$$

- Let $\mathbf{y} = [y^{(1)}, y^{(2)}, \ldots, y^{(N)}]^T$. The ML estimate of $\mathbf{w}$ is:

$$\mathbf{w}_{ML} = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \mathbf{y}$$

Bayesian Linear Regression

- The conjugate prior of $\mathbf{w}$:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0)$$

- The likelihood function:

$$p(\mathbf{y}|\mathbf{x};\mathbf{w}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \mathbf{w}^T\phi(x^{(i)}))^2}{2\sigma^2}\right)$$

- The posterior distribution is then given by $p(\mathbf{w}|\mathbf{y}) = \mathcal{N}(\mathbf{m}_N, \mathbf{S}_N)$ where

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \frac{\boldsymbol{\Phi}^T\mathbf{y}}{\sigma^2}) \quad (3)$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \frac{\boldsymbol{\Phi}^T\boldsymbol{\Phi}}{\sigma^2} \quad (4)$$

Bayesian Linear Regression (2)

- Suppose

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

- The posterior distribution over $\mathbf{w}$ is then given by

$$\mathbf{m}_N = \frac{\mathbf{S}_N \mathbf{\Phi}^T \mathbf{y}}{\sigma^2} \quad (5)$$

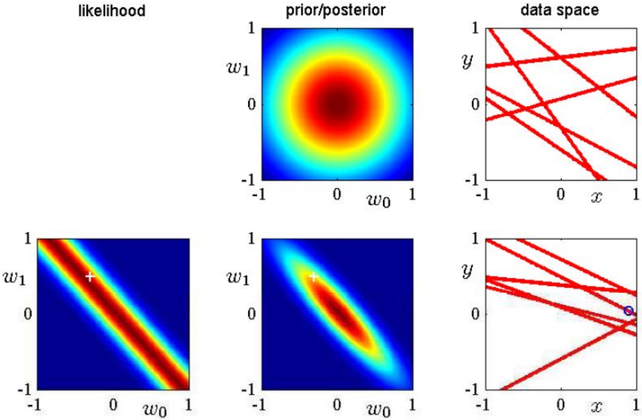
$$\mathbf{S}_N^{-1} = \alpha \mathbf{I} + \frac{\mathbf{\Phi}^T \mathbf{\Phi}}{\sigma^2} \quad (6)$$

- Maximizing the posterior distribution w. r. t. $\mathbf{w}$ is equivalent to minimizing the sum-of-squares error function with a quadratic regularization term, with $\lambda = \alpha/\sigma^2$.

$$\ln p(\mathbf{w}|\mathbf{y}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y^{(i)} - \mathbf{w}^T\phi(x^{(i)}))^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + const$$

Bayesian Linear Regression (3)

- Illustration of sequential Bayesian learning for a simple linear model $f(x) = w_0 + w_1 x$:



Bayesian Linear Regression (4)

- Illustration of sequential Bayesian learning for a simple linear model $f(x) = w_0 + w_1 x$: