6.867: Exercises (Week 3)

Sept 8, 2016

Since there are no recitations this week, here's how we suggest you go through the exercises:

- 2 star exercises (4, 5, 9.1, 9.2, 10, 11) are the ones you should prioritize, to make sure you understand the underlying concepts for regression, discriminant analysis and Naïve Bayes.
- 1 star exercises (6, 12) are also very good practice, and will provide a more in-depth understanding of the material.
- And of course, feel free to take a look at all the others! They either expand on this week's material even further, or go over material covered in the previous weeks.

Contents

1 Two stories **2**

2 Bayesian astronomy **2**

3 Car Bocce **2**

4 Bayesian Regression ** **3**

5 The New Normal ** **5**

6 Where did you put the map? * **5**

7 Posterior predictive distribution **6**

8 Monotony **6**

9 Gaussian decision boundaries ** **6**

10 QDA with 3 classes ** **7**

11 Naive Bayes with Mixed Features ** **8**

12 Probabilistic Models for Classification * **8**

13 Classification Questions from Bishop (note notational differences) **9**

1 Two stories

Which of these stories about Bayesian estimation is correct?

1. There is a random variable V (such as the distance a car will travel after I step on the brakes). We think it has a Gaussian distribution, and that the variance of that distribution is σ^2. We are not sure about the mean M of that distribution, and we model our lack of knowledge about the mean of that distribution using a Gaussian with mean μ_0 and variance σ_0^2. Our data consists of samples of that random variable, which we use to compute a new distribution, parameterized by μ_n and σ_n, on M.

2. There is an actual non-random quantity ν in the world (such as the distance a car will travel) but we don't know it. We model our lack of knowledge of ν using a Gaussian distribution with mean μ_0 and variance σ_0^2. We make measurements of that quantity that are distributed with mean ν and variance σ^2, which we use to update our "belief" about ν in the form of a Gaussian with parameters μ_n and σ_n.

2 Bayesian astronomy

You are making observations of the position of a star, in the 2D coordinate frame of your telescope. You believe you have pointed your scope at the star, and that it generates errors with standard deviation of 10 pixels. The covariance Σ_D of the observations is $((100, 0), (0, 100)$. Your prior on the position of the star is Normal, with mean $\mu = (0, 0)$ and covariance $\Sigma = ((10, 5), (5, 10)$.

1. You make an observation of $(3, 4)$. What is your posterior on the star's position?

2. You make another observation, this time at $(2, 1)$. Now what is the star's position?

3 Car Bocce

This problem needs to be solved numerically.

You just bought a new car and are going to compete in an exciting new extreme sport called "Car Bocce."[1] The idea is to drive up to a wall and stop as close to it as possible without hitting it. The cars in this sport are specially modified so that they move forward at a fixed velocity until the driver pushes a "brake" button, at which point they brake as hard as possible until they come to rest. The driver's job is to select a distance d from the wall at which to push the brake button.

The *braking distance* B of your car is how far it will travel after you have pushed the brake button; it is stochastic, so B is a random variable. You think the braking distance is well modeled

[1] Bocce (also known as boules or petanque) is a game in which one tries to roll balls so that they stop as close as possible to another ball. Our game also resembles a game played by kids in Israel who try to throw apricot pits close to a wall (and penny-pitching, which is the same, but in the US with pennies).

with a Gaussian distribution with $\sigma = 1\text{m}$, but you are uncertain about the mean. You model your belief about the mean with $\mu_0 = 10$ and $\sigma_0^2 = 100$.

The loss in this problem is 1000 if you hit the wall and otherwise $d - b$ where b was the actual braking distance.

(a) What is the optimal distance d^* at which to push the brake button? If you follow that strategy how likely is it that you will crash? What is your risk?

(b) You try this for a few times and (miraculously!) your car remains intact. You update your belief about the mean with the data you gathered that way and so the posterior distribution on the mean of B's distribution is $\mu_5 = 10$ and $\sigma_5^2 = 4^2$.

Now what is the optimal d^*? What is the likelihood of a crash? What is the risk?

4 Bayesian Regression **

In this problem we will consider the standard Bayesian approach to linear regression, in which we put a Gaussian prior on the weights. Assume $x^{(i)} \in \mathbb{R}^2$, where the first feature of each $x^{(i)}$ is 1.

$$\begin{aligned} Y \mid X &\propto \text{Normal}(W^T X, 1) \\ W &\propto \text{Normal}(\mathbf{0}, \mathbf{I}) \end{aligned}$$

The figure below has some plots of $\Pr(W \mid \mathcal{D})$ and $\Pr(\mathcal{D} \mid W)$ for different values of $\mathcal{D}$. Each plot is in the space of weights, indexed by w_1 and w_2, so that the mean of $\Pr(y \mid x_2) = w_1 + w_2 x_2$.

In the densities, the smallest contour contains 10% of the probability mass, and each larger contour is the next decile. In the likelihood plots, the brighter areas have higher density.

When writing the data below we are showing the data points written as (x_2, y) pairs (since the x_1 value for each item is 1).

For each of the following quantities, indicate which plot corresponds to it, or **None** if none of them do.

(a) $\Pr(W)$
◯ A ◯ B ◯ C ◯ D ◯ E ◯ F ◯ G ◯ H ◯ I ◯ None

(b) $\Pr(\mathcal{D} = \{(1,1)\} \mid W)$
◯ A ◯ B ◯ C ◯ D ◯ E ◯ F ◯ G ◯ H ◯ I ◯ None

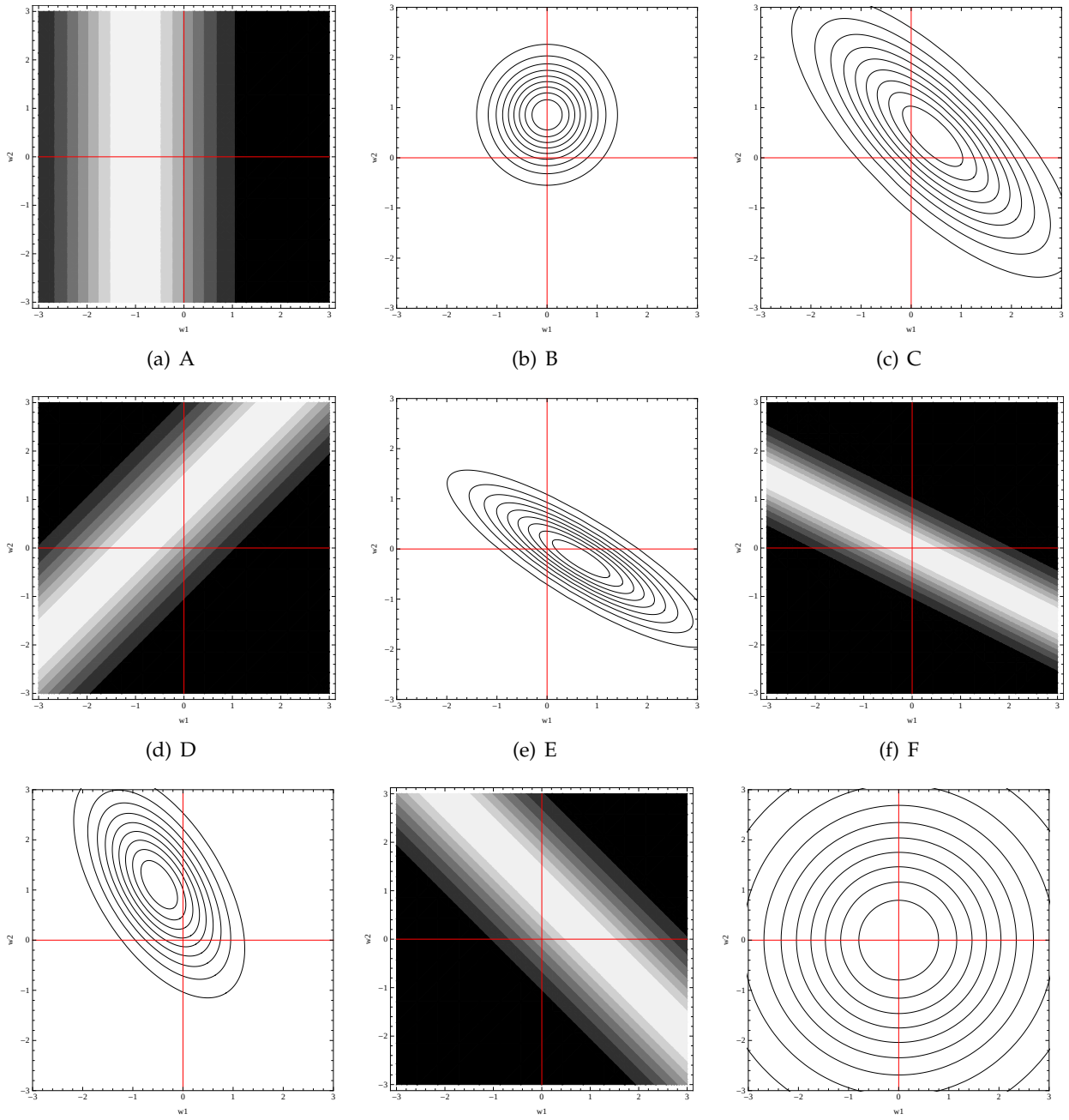
(c) $\Pr(\mathcal{D} = \{(-1,-1)\} \mid W)$
◯ A ◯ B ◯ C ◯ D ◯ E ◯ F ◯ G ◯ H ◯ I ◯ None

(d) $\Pr(\mathcal{D} = \{(0,-1)\} \mid W)$
◯ A ◯ B ◯ C ◯ D ◯ E ◯ F ◯ G ◯ H ◯ I ◯ None

(e) $\Pr(W \mid \mathcal{D} = \{(1,1)\})$
◯ A ◯ B ◯ C ◯ D ◯ E ◯ F ◯ G ◯ H ◯ I ◯ None

(f) $\Pr(W \mid \mathcal{D} = \{(1,1),(-1,-1)\})$
◯ A ◯ B ◯ C ◯ D ◯ E ◯ F ◯ G ◯ H ◯ I ◯ None

(g) $\Pr(W \mid \mathcal{D} = \{(1,1),(0,-1)\})$
◯ A ◯ B ◯ C ◯ D ◯ E ◯ F ◯ G ◯ H ◯ I ◯ None



(a) A (b) B (c) C

(d) D (e) E (f) F

(g) G (h) H (i) I

Figure 1: **Linear Regression Plots**

5 The New Normal **

1. You're an hour late for 6.867 lecture and walk in to see the following formulas written on the board; the lecture seems to have been about a Bayesian approach to linear regression. There are several symbols that seem to denote some form of variance. Pick which of the following explanations go with each one, or argue that it doesn't apply.

$$\begin{aligned} p(\theta) &= \mathcal{N}(\theta \mid \mu_0, \Sigma_0) \\ p(y \mid x, \theta) &= \mathcal{N}(\theta^T x, \sigma^2) \\ p(\theta \mid \mathcal{D}) &= \mathcal{N}(\theta \mid \mu_n, \Sigma_n) \\ \mu_n &= \Sigma_n(\Sigma_0^{-1}\mu_0 + (1/\sigma^2)\Phi^T Y) \\ \Sigma_N^{-1} &= \Sigma_0^{-1} + \frac{1}{\sigma^2}\Phi^T\Phi \\ p(y^{(n+1)} \mid x^{(n+1)}, \mathcal{D}) &= \mathcal{N}(y \mid \mu_n^T x^{(n+1)}, x^{(n+1)^T}\Sigma_N x^{(n+1)} + \sigma^2) \end{aligned}$$

   (a) Variance of prior prediction.
   (b) Variance of posterior prediction.
   (c) Prior variance on mean of weight distribution.
   (d) Variance of measurements.
   (e) Posterior variance on mean of weight distribution.

2. You have just discovered a time machine and want to use it to regress back in time to your first birthday. There is a big knob that seems to be freely turnable in both directions; when you turn it, there is a numeric "read-out" on the console of the time machine that varies linearly with the amount the knob is turned. Right now, the numbers on the display read 2016.75, which happens to be the current time, measured in years. You think that the amount the knob is turned correlates with how far forward or backward in time the machine goes.

   You begin to do some experiments. You find that when you arrive at a new time, you can estimate the year, with a standard deviation of about 2 years. Your best guess, initially, is that the the display is in direct correspondence with the date, but you assign a variance of 1 to the parameters of the linear dependence and you don't think the parameters are correlated.

   (a) You turn the knob to 2000. What is your distribution on what year you will end up in?
   (b) Once there, you realize that the year is 1015. What is your distribution on the parameters governing the relationship between the knob and the year?
   (c) You turn the knob to 2010. What is your distribution on the year you will end up in?
   (d) What are the numeric values of the variances listed at the end of the previous question?

6 Where did you put the map? *

In class, we derived the ridge regression estimator for linear regression with a prior by taking the derivative, setting it to 0, and solving. An alternative is to observe that the posterior distribution is a Gaussian and is, therefore unimodal and symmetric. Consequently, its mean is also its mode.

What is its mode? How is it related to the ridge-regression result?

7 Posterior predictive distribution

We know this went by in lecture, but it was probably kind of fast. Derive the formula for the posterior predictive distribution for linear regression with a Gaussian prior on the weights.

8 Monotony

In the basic Gaussian parameter-estimation case (not regression), we find that after getting one data sample, the new precision matrix is the sum of the old precision matrix and the precision of the observation:

$$\Sigma_1^{-1} = \Sigma_0^{-1} + \Sigma^{-1}$$

where Σ_0 is the prior covariance on the mean of the distribution and Σ is the known covariance.

In lecture we said informally that "the variance decreases monotonically" as a the amount of data increases. We'll try to make it more formal and call it a theorem. Here are some possible theorems, some harder to prove than others. Try to prove them. Don't worry if you don't get them all. You'll probably find the *Matrix Cookbook*[2] handy.

(a) **Theorem 1** The trace of Σ_n^{-1} increases monotonically as n increases.

(b) **Theorem 2** The minimum eigenvalue of Σ_n decreases monotonically as n increases.

(c) **Theorem 3** The trace of Σ_n decreases monotonically as n increases.

> You can use a special case of the Woodbury identity: for invertible matrices of the correct sizes,
>
> $$(A + B)^{-1} = A^{-1} - A^{-1}(A^{-1} + B^{-1})^{-1}A^{-1}.$$

9 Gaussian decision boundaries **

1. Murphy 4.21

   Let $p(x|y = j) = \mathcal{N}(x|\mu_j, \sigma_j)$ where $j = 1, 2$ and $\mu_1 = 0, \sigma_1^2 = 1, \mu_2 = 1, \sigma_2^2 = 10^6$.
   Let the class priors be equal, $p(y = 1) = p(y = 2) = 0.5$

   (a) Find the decision region

   $$R_1 = \{x : p(x|\mu_1, \sigma_1) \geqslant p(x|\mu_2, \sigma_2)\}$$

   Sketch the result. Hint: draw the curves and find where they intersect by solving $p(x|\mu_1, \sigma_1) = p(x|\mu_2, \sigma_2)$.

   (b) Now suppose $\sigma_2 = 1$ (and all other parameters remain the same). what is R_1 in this case?

2. Decision boundary types

   Use these functions to investigate different possible Gaussian decision boundaries. For each of the following specifications, find an appropriate pair of Gaussians and prior class probabilities.

[2] http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3274

(a) A linear decision boundary.

(b) A linear decision boundary, with both means on the same side of the decision boundary.

(c) A parabolic decision boundary.

(d) A non-continuous decision boundary (one class represented by 2 disconnected regions).

(e) A circular decision boundary.

(f) A skewed ellipsoid decision boundary, with only one mean inside the ellipsoid.

(g) No decision boundary; the entire plane is one decision region.

3. In many cases, it is necessary to classify into more than two classes. A natural extension of the Gaussian mixture approach is to fit a Gaussian distribution for each class, and classify each input vector to the class with the highest posterior probability for it.

   We would like to modify the function `plot_two_gauss2d` to plot the decision boundaries between three Gaussians. Write a new `plot3gaussians` procedure and use it to plot decision boundaries on several examples.

   (a) All decision boundaries are linear.

   (b) Some decision boundaries are linear, while others are quadratic.

   (c) All decision boundaries are quadratic.

   (d) There are only two decision regions (one class never gets selected).

   ALSO: mark each decision region in the plots with an appropriate label.

10 QDA with 3 classes **

Consider a three category classification problem. Let the prior probabilities be

$$P(Y = 1) = P(Y = 2) = P(Y = 3) = \frac{1}{3}.$$

The class-conditional densities are multivariate normal densities with parameters

$$\mu_1 = [0, 0]^T, \qquad \mu_2 = [1, 1]^T, \qquad \mu_3 = [-1, 1]^T$$

$$\Sigma_1 = \begin{bmatrix} 0.7 & 0 \\ 0 & 0.7 \end{bmatrix}, \qquad \Sigma_2 = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}, \qquad \Sigma_3 = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$$

Classify the following points:

(a) x = [-0.5, 0.5]

(b) x = [0.5, 0.5]

11 Naive Bayes with Mixed Features **

Consider a three-class naive Bayes classifier with one binary feature, one Gaussian feature and multinomial output:

$$y \sim \text{Mu}(y \mid \pi, 1), \qquad x_1 \mid y = c \sim \text{Ber}(x_1 \mid \theta_c), \qquad x_2 \mid y = x \sim \mathcal{N}(x_2 \mid \mu_c, \sigma_c^2)$$

where the subscript c denotes the class.

Let the parameter vectors be as follow:

$$\pi = (0.5, 0.25, 0.25), \qquad \theta = (0.5, 0.5, 0.5), \qquad \mu = (-1, 0, 1), \qquad \sigma^2 = (1, 1, 1)$$

(a) Compute $p(y \mid x_1 = 0, x_2 = 0)$ (the result should be a vector of 3 numbers that sums to 1).

(b) Compute $p(y \mid x_2 = 0)$.

(c) Compute $p(y \mid x_1 = 0)$.

(d) Explain any interesting patterns you see in your results.

12 Probabilistic Models for Classification *

In this question we'll consider feature selection in a naive Bayes model. As in the lecture, we'll assume that the training set consists of examples $(\mathbf{x}^{(i)}, y^{(i)})$ where labels $y^{(i)} \in \{-1, +1\}$ and features $x_i \in \{0, 1\}^d$. In a regular naive Bayes model, we define

$$p(\mathbf{x}, y; \theta) = p(y) \prod_{j=1}^{d} p_j(x_j \mid y) \tag{12.1}$$

Now consider a naive Bayes model defined on a subset of all possible features. We use A to define the set of active features in the model. The set A is a subset of $\{1, 2, \ldots, d\}$. The model then takes the following form:

$$p(\mathbf{x}, y; \theta, A) = p(y) \prod_{j \in A} p_j(x_j \mid y) \prod_{j \notin A} p_j(x_j) \tag{12.2}$$

Note that for each active feature we have parameters of the form $p_j(x \mid y)$ which depend on the label y, and for inactive features we have parameters of the form $p_j(x)$ which ignore the label y.

The $p_j(x_j)$ terms are necessary to ensure that the model still correctly defines a distribution $p(\mathbf{x}, y)$ over all possible $\mathbf{x}, y$ pairs. It might be tempting to write the model as

$$p(\mathbf{x}, y; \theta, A) = p(y) \prod_{j \in A} p_j(x_j \mid y)$$

but under this definition the x_j terms for $j \notin A$ are not accounted for in the model, and the model is deficient (in fact, it can be shown that in this case $\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(\mathbf{x}, y; \theta, A) > 1$, assuming that $\mathcal{X} = \{0, 1\}^d$ and $\mathcal{Y} = \{-1, +1\}$, clearly violating the laws of probability).

1. The classification function is

$$f(\mathbf{x}) = \arg\max_y p(\mathbf{x}, y; \theta, A) = \text{sign}[\log p(\mathbf{x}, +1; \theta, A) - \log p(\mathbf{x}, -1; \theta, A)]$$

Show that for any feature $j \notin A$, its value is irrelevant to the classification function. For example, for any feature vector $\mathbf{x}$, if we defined a new feature vector $\mathbf{x}'$ where $x'_j = x_j$ for $j \in A$, and $x'_j = 1 - x_j$ for $j \notin A$, then $f(\mathbf{x}) = f(\mathbf{x}')$.

2. Now let's consider a feature-selection method for these models. In this method, initially we start with A equal to the empty set; i.e., there are no active features. At each iteration we greedily choose a single feature which gives the most "improvement" in the model. We will define a precise measure of "improvement" below. Assume that maximum-likelihood estimates are used in the model:

$$\hat{p}_j(\mathbf{x} \mid y) = \frac{\sum_{i=1}^n \text{count}[x_j^{(i)} = x \text{ and } y^{(i)} = y]}{\sum_{i=1}^n \text{count}[y^{(i)} = y]}$$

and

$$\hat{p}_j(\mathbf{x}) = \frac{\sum_{i=1}^n \text{count}[x_j^{(i)} = x]}{n} \tag{12.3}$$

We measure the gain for any feature $k \notin A$ given a set of active features A as

$$\text{Gain}(k; A) = \sum_i \log p(\mathbf{x}^{(i)}, y^{(i)}; \theta', A \cup \{k\}) - \sum_i \log p(\mathbf{x}^{(i)}, y^{(i)}; \theta', A) \tag{12.4}$$

$\text{Gain}(k; A)$ simply measures how much adding the k^{th} feature improves the fit of the model to the training data, where "fit of the model" is measured as log-likelihood. At each point in the feature selection method, we choose the feature with the maximal gain. At some point, we stop adding features (for example, we might use cross-validation to choose the stopping point), so that the final model uses a relatively small subset of the features.

Show that

$$\text{Gain}(k; A) = n \sum_{y \in \{-1,+1\}} \sum_{x \in \{0,1\}} \hat{p}_k(x, y) \log \frac{\hat{p}(x \mid y)}{\hat{p}_k(x)} \tag{12.5}$$

where

$$\hat{p}_k(x, y) = \frac{1}{n} \sum_{i=1}^n \text{count}[x_k^{(i)} = x \text{ and } y^{(i)} = y] \tag{12.6}$$

$\text{Gain}(k; A)$ is an estimate, based on the training examples, of the mutual information between the feature x_k and the label y. Mutual information is a quantity that measures the dependence between two random variables (in this case X_k and Y), and that arises frequently in probability theory and information theory.

13 Classification Questions from Bishop (note notational differences)

1. Bishop 4.9

Consider a generative classification model for K classes defined by prior class probabilities $p(C_k) = \pi_k$ and general class-conditional densities $p(\phi|C_K)$ where ϕ is the input feature vector. Suppose we are given a training data set $\{\phi_n, t_n\}$ where $n = 1, \ldots, N$, and t_n is a binary target vector of length K that uses the 1-of-K coding scheme, so that it has components $t_{nj} = I_{jk}$ if pattern n is from class C_k. Assuming that the data points are drawn independently from this model, show that the maximum-likelihood solution for the prior probabilities is given by

$$\pi_k = \frac{N_k}{N} \tag{13.1}$$

where N_k is the number of data points assigned to class C_k.

2. Bishop 4.10

Consider the classification model of Exercise 4.9 (above) and now suppose that the class-conditional densities are given by Gaussian distributions with a shared covariance matrix:

$$p(\phi|C_k) = \mathcal{N}(\phi|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \tag{13.2}$$

Show that the maximum likelihood solution for the mean of the Gaussian distribution for class C_k is given by

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} \phi_n \tag{13.3}$$

which represents the mean of those feature vectors assigned to class C_k. Similarly, show that the maximum likelihood solution for the shared covariance matrix is given by

$$\boldsymbol{\Sigma} = \sum_{k=1}^{K} \frac{N_k}{N} \mathbf{S}_k \tag{13.4}$$

where

$$\mathbf{S}_k = \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} (\phi_n - \boldsymbol{\mu}_k)(\phi_n - \boldsymbol{\mu}_k)^T \tag{13.5}$$

Thus $\boldsymbol{\Sigma}$ is given by a weighted average of the covariances of the data associated with each class, in which the weighting coefficients are given by the prior probabilities of the classes.

3. Bishop 4.11 - Naïve Bayes

Consider a classification problem with K classes for which the feature vector ϕ has M components each of which can take L discrete states. Let the values of the components be represented by a 1-of-L binary coding scheme. Further suppose that, conditioned on the class C_k, the M components of ϕ are independent, so that the class-conditional density factorizes with respect to the feature vector components. Show that the quantities a_k given by

$$a_k = \ln(p(\mathbf{x}|C_k)p(C_k)) \tag{13.6}$$

which appear in the argument to the softmax function describing the posterior class probabilities, are linear functions of the components of ϕ. Note that this represents an example of the naive Bayes model which is discussed in Bishop.