

**COMP 6467/4670 Introduction to Statistical Machine Learning
Tutorial 1 - Solution Sheet**

March 21, 2008

Problem 1 (Regression - ϵ -insensitive Loss)

Assume we have the following loss function

$$c(\mathbf{x}, y, f(\mathbf{x})) = |y - f(\mathbf{x})|_\epsilon \text{ where } |\xi|_\epsilon := \begin{cases} 0 & \text{if } |\xi| \leq \epsilon \\ \xi - \epsilon & \text{if } \xi > \epsilon \\ -\xi - \epsilon & \text{otherwise} \end{cases}$$

1. Rewrite $|\xi|_\epsilon$ as a linear optimization problem. **Hint:** all you need to do is to modify the constraints.
2. Rewrite the regularized risk functional for a linear model $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$ as a quadratic optimization problem with constraints. **Hint:** you only need to take care of the empirical risk term. Recall that the regularized risk is given by

$$R_{\text{reg}}[f] = \frac{1}{m} \sum_{i=1}^m |y_i - f(\mathbf{x}_i)|_\epsilon + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

Solution 1

1. Given the following loss function

$$c(\mathbf{x}, y, f(\mathbf{x})) = |y - f(\mathbf{x})|_\epsilon = |\xi|_\epsilon = \max(0, |\xi| - \epsilon).$$

Transferring this minimization of loss function to a constrained linear programming can be done by introducing new variables $\nu \geq 0$, $\nu^* \geq 0$ that *upper bound* the loss so that $|\xi|_\epsilon = \nu + \nu^*$ for the minimal ν and ν^* :

$$\begin{aligned} & \text{minimize} && \nu + \nu^* \\ & \text{subject to} && \begin{cases} \xi - \epsilon \leq \nu \\ -\xi - \epsilon \leq \nu^* \\ \nu, \nu^* \geq 0 \end{cases} \end{aligned} \tag{1}$$

Note that either ν or ν^* will be 0 for any given ξ .

2. We are interested to solve the following optimization problem

$$\begin{aligned} & \underset{f}{\text{minimize}} && R_{\text{reg}}[f] \\ & = \underset{f}{\text{minimize}} && \frac{1}{m} \sum_{i=1}^m |y_i - f(\mathbf{x}_i)|_\epsilon + \frac{\lambda}{2} \|\mathbf{w}\|^2. \end{aligned}$$

From (1) and for a class of linear function f , the above optimization problem would be

$$\begin{aligned} & \underset{\mathbf{w}, \nu, \nu^*}{\text{minimize}} && \frac{1}{m} \sum_{i=1}^m (\nu_i + \nu_i^*) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \\ & \text{subject to} && \begin{cases} y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - \epsilon \leq \nu_i \\ \langle \mathbf{w}, \mathbf{x}_i \rangle - y_i - \epsilon \leq \nu_i^* \\ \nu_i, \nu_i^* \geq 0 \end{cases} \end{aligned} \quad (2)$$

Problem 2 (Classification - Linear and Quadratic Discriminant Function)

A 2-class Bayes classifier on \mathbb{R}^2 has equal a priori probabilities and normal conditional densities.

1. Find its discriminant functions and its decision boundary for the following means and covariances

$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \Sigma_1 = \Sigma_2 = \Sigma = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 2 \end{bmatrix}$$

2. Find its discriminant functions and its decision boundary for the following means and covariances

$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \Sigma_1 = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 2 \end{bmatrix}; \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

Solution 2

1. Given equal a priori probabilities, i.e. $p(y = 1) = p(y = 2) = 0.5$. We note the equation for n-dimensional normal conditional density is

$$p(\mathbf{x}|y = i) = \frac{1}{(2\pi)^{n/2} |\Sigma_i|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) \right). \quad (3)$$

In our problem, $n = 2$. The Bayes discriminant function for class-1 is given by

$$\begin{aligned} g_1(\mathbf{x}) &= \ln p(y = 1|\mathbf{x}) \\ &\propto \ln p(\mathbf{x}|y = 1) + \ln p(y = 1) \\ &\propto -\ln(2\pi \times \sqrt{0.5 \times 2}) - \frac{1}{2} \left(2(x_1 - 3)^2 + \frac{1}{2}(x_2 - 6)^2 \right) + \ln(0.5). \end{aligned}$$

For class-2, the discriminant function is

$$\begin{aligned} g_2(\mathbf{x}) &= \ln p(y = 2|\mathbf{x}) \\ &\propto \ln p(\mathbf{x}|y = 2) + \ln p(y = 2) \\ &\propto -\ln(2\pi \times \sqrt{0.5 \times 2}) - \frac{1}{2} \left(2(x_1 - 3)^2 + \frac{1}{2}(x_2 + 2)^2 \right) + \ln(0.5). \end{aligned}$$

Since the classifier assigns a feature vector \mathbf{x} to class- i if

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \neq i, \quad (4)$$

the decision boundary of the above problem could be found by setting

$$\begin{aligned} g_1(\mathbf{x}) &= g_2(\mathbf{x}) \\ \frac{1}{2} \left(2(x_1 - 3)^2 + \frac{1}{2}(x_2 - 6)^2 \right) &= \frac{1}{2} \left(2(x_1 - 3)^2 + \frac{1}{2}(x_2 + 2)^2 \right). \end{aligned}$$

Solving the above equation, will give us $x_2 = 2$ as the decision boundary. It is a *linear* decision boundary. (Question: this comparison requires that the normalizing constants for both g_1 and g_2 are the same. Why does this hold in this case?)

2. For in-equal covariance matrices, the discriminant function for class-1 is given by

$$\begin{aligned} g_1(\mathbf{x}) &= \ln p(y = 1|\mathbf{x}) \\ &\propto \ln p(\mathbf{x}|y = 1) + \ln p(y = 1) \\ &\propto -\ln(2\pi \times \sqrt{0.5 \times 2}) - \frac{1}{2} \left(2(x_1 - 3)^2 + \frac{1}{2}(x_2 - 6)^2 \right) + \ln(0.5), \end{aligned} \quad (5)$$

and for class-2 by

$$\begin{aligned} g_2(\mathbf{x}) &= \ln p(y = 2|\mathbf{x}) \\ &\propto \ln p(\mathbf{x}|y = 2) + \ln p(y = 2) \\ &\propto -\ln(2\pi \times \sqrt{2 \times 2}) - \frac{1}{2} \left(\frac{1}{2}(x_1 - 3)^2 + \frac{1}{2}(x_2 + 2)^2 \right) + \ln(0.5). \end{aligned} \quad (6)$$

Similarly, to find the decision boundary, we equate

$$g_1(\mathbf{x}) = g_2(\mathbf{x}).$$

Its decision boundary is then $x_2 = 0.1875x_1^2 - 1.125x_1 + 3.514$. It is a *quadratic* decision boundary.

This shows that modeling discriminant function with equal covariances produces a linear decision function, whereas with different covariances produces a quadratic decision function.

Problem 3 (Classification - Decision Tree)

You are trying to learn a strategy for Texas Hold'em using Decision Trees. For your two received cards you define 3 attributes:

- Have Ace : is one of the cards an Ace? Can be either yes or no.
- Same Suit: are the two cards of the same suit? Can be either yes or no.
- Difference: how close are the two cards by value? Can be either 0,1 or 2, or > 2 .

The task is to learn whether to fold or not. Construct a Decision Tree based on the following 10 observations. If there is more than one attribute that gives the best split then pick the one that comes earlier alphabetically.

Have Ace	Same Suit	Difference	Fold
no	yes	> 2	yes
no	yes	1 or 2	no
no	no	> 2	yes
yes	no	0	no
no	no	1 or 2	yes
yes	yes	> 2	yes
yes	no	1 or 2	no
yes	yes	1 or 2	no
no	no	0	no
yes	no	> 2	yes

Solution 3 We note the definition of information entropy of a discrete random variable $X = \{x_1, x_2, x_3, \dots, x_n\}$ as

$$I = - \sum_{i=1}^n p(x_i) \log_2(p(x_i)),$$

where $p(x_i)$ denotes probability mass function of X .

Iteration 1:

	Fold		
Have Ace	yes	no	I
yes	2	3	0.971
no	3	2	0.971
			$\sum = 0.971$

	Fold		
Same Suit	yes	no	I
yes	2	2	1
no	3	3	1
			$\sum = 1$

Difference \ Fold	yes	no	I
	0	0	2
1 or 2	1	3	0.81
> 2	4	0	0
			$\sum = 0.324$

Split on lowest information entropy attribute, i.e. Difference.

Iteration 2:

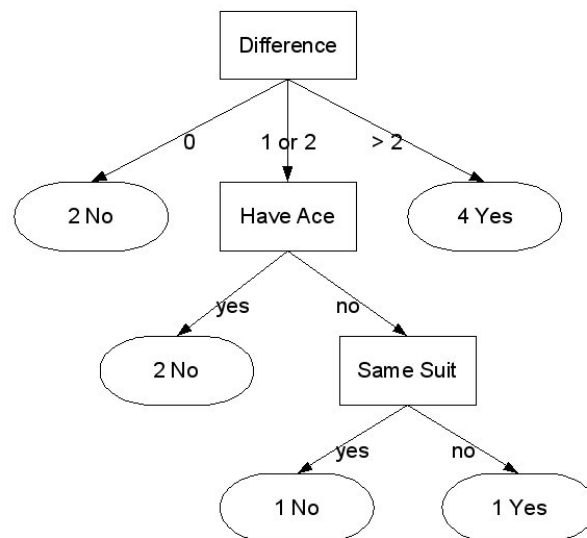
Have Ace \ Fold	yes	no	I
	yes	0	2
no	1	1	1
			$\sum = 0.5$

Same Suit \ Fold	yes	no	I
	yes	0	2
no	1	1	1
			$\sum = 0.5$

Since the two attributes have equal entropy value, we split on Have Ace because it comes earlier alphabetically.

Iteration 3: Split on the only remaining attribute, i.e. Same Suit.

The above decision tree generation procedure is known as *Iterative Dichotomiser 3* or ID3. The resulting decision tree is as follows:



Extra 1 (Derivation of Ridge Regression Estimation)

We are interested to solve the following optimization problem

$$\begin{aligned} & \underset{w}{\text{minimize}} && E'(w) = \\ & \underset{w}{\text{minimize}} && E(w) + R(w) \\ & && \text{where :} \\ & && E(w) = \|y - Xw\|_2^2 \text{ and} \\ & && R(w) = \lambda \|w\|_2^2. \end{aligned}$$

Since the above regularized loss is convex in w (its Hessian matrix is positive semi-definite), we just need to satisfy the first order optimality condition, which is

$$\begin{aligned} (\partial/\partial w)E'(w) &= 0 \\ (\partial/\partial w) ((y - Xw^*)^T(y - Xw^*) + \lambda w^{*T}w^*) &= 0 \\ -2X^T y + 2X^T Xw^* + 2\lambda w^* &= 0 \\ w^* &= (X^T X + \lambda I)^{-1} X^T y. \end{aligned} \tag{7}$$

Extra 2 (Derivation of Bias-Variance Tradeoff)

Note the MSE formulation

$$MSE(x_0) = \mathbb{E}_X[(f_X(x_0) - y_0)^2], \tag{8}$$

the trick is to augment the following

$$f_{\text{ave}}(x_0) := \mathbb{E}_X[f_X(x_0)], \tag{9}$$

to the above equation. Thus,

$$\begin{aligned} MSE(x_0) &= \mathbb{E}_X[((f_X(x_0) - f_{\text{ave}}(x_0)) + (f_{\text{ave}}(x_0) - y_0))^2] \\ &= \mathbb{E}_X[(f_X(x_0) - f_{\text{ave}}(x_0))^2 + (f_{\text{ave}}(x_0) - y_0)^2 + 2(f_X(x_0) - f_{\text{ave}}(x_0))(f_{\text{ave}}(x_0) - y_0)] \\ &\stackrel{(a)}{=} \mathbb{E}_X[(f_X(x_0) - f_{\text{ave}}(x_0))^2] + (f_{\text{ave}}(x_0) - y_0)^2 + 2(f_{\text{ave}}(x_0) - y_0)\mathbb{E}_X[f_X(x_0) - f_{\text{ave}}(x_0)] \\ &\stackrel{(b)}{=} \mathbb{E}_X[(f_X(x_0) - f_{\text{ave}}(x_0))^2] + (f_{\text{ave}}(x_0) - y_0)^2 \\ &= \text{variance} + \text{bias}^2. \end{aligned}$$

In (a), we have used linearity of expectation operator and noticed that the term $(f_{\text{ave}}(x_0) - y_0)$ is constant with respect to expectation. For (b), the term $2(f_{\text{ave}}(x_0) - y_0)\mathbb{E}_X[f_X(x_0) - f_{\text{ave}}(x_0)]$ disappears as $\mathbb{E}_X[f_X(x_0)] = f_{\text{ave}}(x_0)$.