Statistical Machine Learning HW2

1. (a) $H(X,Y) = H(Y|X) + H(X)$

$$H(X,Y) - H(X|Y) - H(Y|X)$$
$$= H(Y|X) + H(X) - H(X|Y) - H(Y|X)$$
$$= H(X) - H(X|Y) = IG(X;Y)$$

(b) (i) Initial Entropy of Usage is

$n = 15$
$L = 7$
$M = 5$
$H = 3$

$$H(S) = -P(Low)\cdot\log_2(P(Low)) - P(Medium)\cdot\log_2(P(Med)) - P(High)\cdot\log_2(P(High))$$
$$= -(7/15)\cdot\log_2(7/15) - (5/15)\cdot\log_2(5/15) - (3/15)\cdot\log_2(3/15)$$
$$= 1.5058$$

(ii) I want to choose the attribute which yields the maximum information gain

First Attribute — Income

categorical values —

	Low			Medium			High		
	5			6			4		
	L	M	H	L	M	H	L	M	H
	5	0	0	2	4	0	0	1	3

$$H(Income: Low) = -(5/5)\cdot\log_2(5/5) - 0 - 0 = 0$$
$$H(Income = Medium) = -(2/6)\cdot\log_2(2/6) - (4/6)\cdot\log_2(4/6) - 0 = 0.9183$$
$$H(Income = High) = -0 - (1/4)\cdot\log_2(1/4) - (3/4)\cdot\log_2(3/4) = 0.81127$$

Average Entropy information for Income.

$$H(Usage|Income) = P(Low)\cdot H(Income = Low) + P(Med)\cdot H(Income = Med) + P(High)\cdot H(Income = High)$$
$$= \frac{5}{15}\cdot 0 + \frac{6}{15}\cdot 0.9183 + \frac{4}{15}\cdot 0.81127$$
$$= 0.58365$$, Information gain $= 1.5058 - 0.58365 = \boxed{0.92215}$

Second Attribute Age

categorical values —

	Old			Young		
	9			6		
	L	M	H	L	M	H
				0	5	1

7 0 2 0 0 1

$H(\text{Age} = \text{old}) = -(7/9) \cdot \log_2(7/9) - 0 - (2/9) \cdot \log_2(2/9) = 0.7642$

$H(\text{Age} = \text{Young}) = -0 - (5/6) \cdot \log_2(5/6) - (1/6) \cdot \log_2(1/6) = 0.65$

Average entropy Information for Age

$$H(\text{Usage} \mid \text{Age}) = P(\text{old}) \cdot H(\text{Age} = \text{old}) + P(\text{young}) \cdot H(\text{Age} = \text{young})$$

$$= (9/15) \cdot 0.7642 + (6/15) \cdot 0.65$$

$$= 0.71852$$

$$\text{Information Gain} = H(S) - H(\text{Usage} \mid \text{Age})$$

$$= 1.5058 - 0.71852$$

$$= \boxed{0.78728}$$

Third Attribute - Education

Categorical values - University, College, High School

University			College			High School		
6			5			4		
L	M	H	L	M	H	L	M	H
3	0	3	0	5	0	4	0	0

$H(\text{Edu} = \text{Univ}) = -(3/6) \cdot \log_2(3/6) - 0 - (3/6) \cdot \log_2(3/6) = 1$

$H(\text{Edu} = \text{College}) = -0 - (5/5) \cdot \log_2(5/5) - 0 = 0$

$H(\text{Edu} = \text{Highschool}) = -(4/4) \cdot \log_2(4/4) - 0 - 0 = 0$

Average Entropy Information for Education

$$H(\text{Usage} \mid \text{Edu}) = P(\text{Univ}) \cdot H(\text{Edu} = \text{Univ}) + P(\text{College}) \cdot H(\text{Edu} = \text{College}) + P(\text{High}) \cdot H(\text{Edu} = \text{High})$$

$$= 6/15 \cdot 1 + 5/15 \cdot 0 + 4/15 \cdot 0$$

$$= 6/15 = 0.4$$

$$\text{Information Gain} = H(S) - H(\text{Usage} \mid \text{Edu})$$

$$= 1.5058 - 0.4$$

$$= \boxed{1.1058}$$

Fourth Attribute - Marital Status

Categorical values – Single Married

	Single			Married		
	7			8		
	L	M	H	L	M	H
	2	2	3	5	3	0

$$H(\text{Marital} = \text{Single}) = -(2/7)\cdot\log_2(2/7) - (2/7)\cdot\log_2(2/7) - (3/7)\cdot\log_2(3/7) = 1.55665$$

$$H(\text{Marital} = \text{Married}) = -(5/8)\cdot\log_2(5/8) - (3/8)\cdot\log_2(3/8) - 0 = 0.95443.$$

$$H(\text{Usage}|\text{Marital}) = P(\text{Single})\, H(\text{Marital} = \text{Single}) + P(\text{Married})\cdot H(\text{Marital} = \text{married})$$

$$= 7/15 \cdot 1.55665 + 8/15 \cdot 0.95443$$

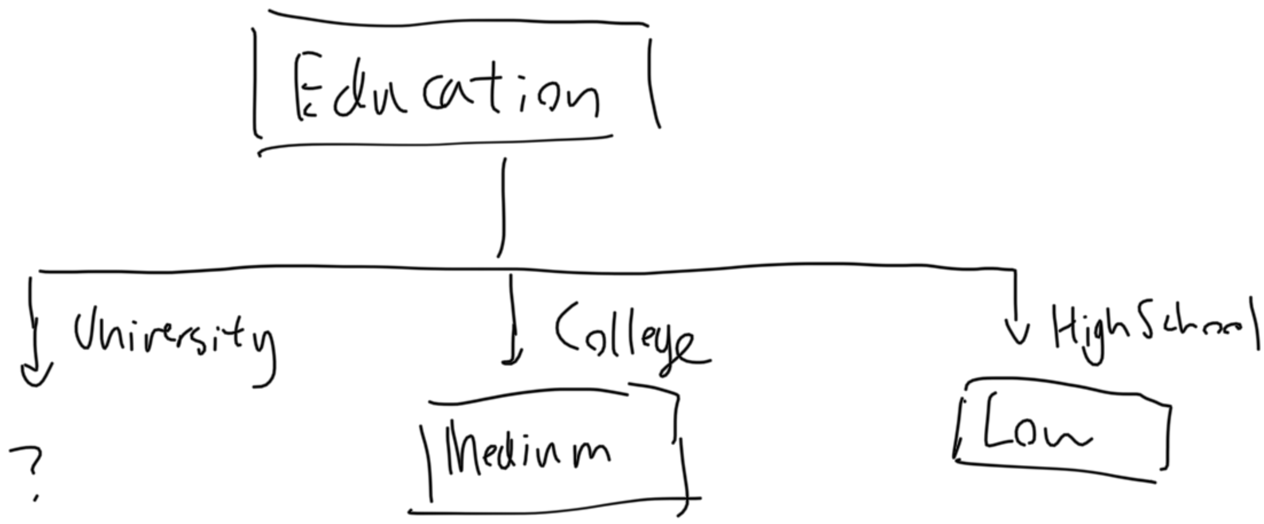
$$= 1.23546$$

$$\text{Information Gain} = H(S) - H(\text{Usage}|\text{Marital})$$

$$= 1.5058 - 1.23546$$

$$= \boxed{0.27034}$$

Here, the attribute with the maximum information gain is Education



Here, when education = college, It's a pure class of medium Usage. when education = High school, It's a pure class of Low usage.. The only thing left is university

Complete entropy of university is.

$$H(S) = -(3/6)\cdot\log_2(3/6) - 0 - (3/6)\cdot\log_2(3/6)$$

$$= 1$$

First Attribute. - Income.

Categorical values. - Low Medium High

	Low			Medium	High		
	3			0	3		
	L	M	H		L	M	H

3 0 0 0 0 3

$H(\text{Univ}, \text{Income} = \text{Low}) = -(3/3)\cdot\log_2(3/3) - 0 - 0 = 0$

$H(\text{Univ}, \text{Income} = \text{Med}) = -0 - 0 - 0 = 0$

$H(\text{Univ}, \text{Income} = \text{High}) = -0 - 0 - (3/3)\log_2(3/3) = 0$

$I(\text{Univ}, \text{Income}) = 0$

Information Gain $= H(\text{Univ}) - I(\text{Univ}, \text{Income}) = 1$

Second Attribute, Age

categories — Old, Young

	Old			Young		
	5			1		
	L	M	H	L	M	H
	3	0	2	0	0	1

$H(\text{Univ}, \text{age} = \text{old}) = -(3/5)\cdot\log_2(3/5) - (2/5)\cdot\log_2(2/5) - 0 = 0.971$

$H(\text{Univ}, \text{age} = \text{young}) = -0 - 0 - (1/1)\cdot\log_2(1/1) = 0$

$I(\text{Univ}, \text{age}) = 5/6 \cdot 0.971 = 0.80916$

Information Gain $= H(\text{Univ}) - I(\text{Univ}, \text{age}) = 0.19084$

Third Attribute — Marital Status.

categories — Single, Married

	Single			Married		
	3			3		
	L	M	H	L	M	H
	0	0	3	3	0	0

$H(\text{Univ}, \text{marital status} = \text{Single}) = -0 - 0 - (3/3)\cdot\log_2(3/3) = 0$

$H(\text{Univ}, \text{marital status} = \text{married}) = -0 - 0 - (3/3)\cdot\log_2(3/3) = 0$

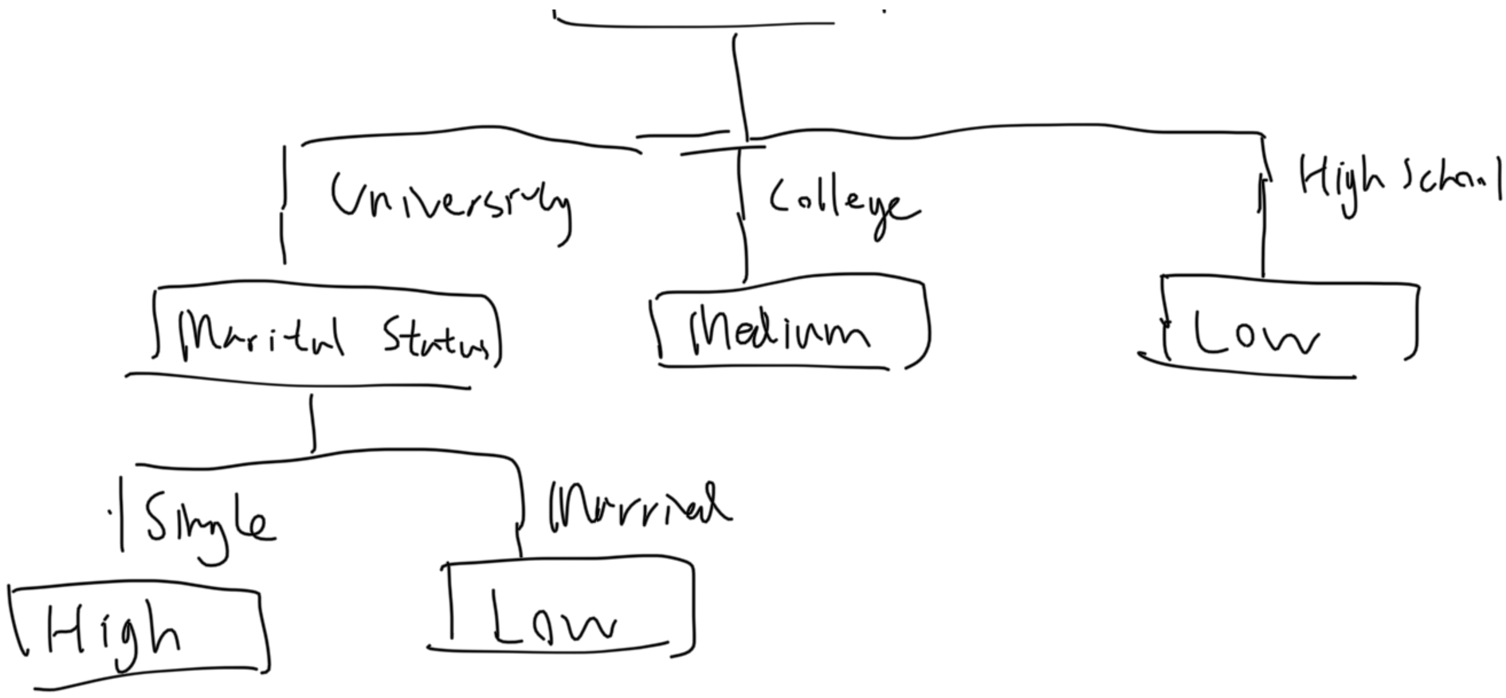
$I(\text{Univ}, \text{marital status}) = 0$

Information Gain $= 1$

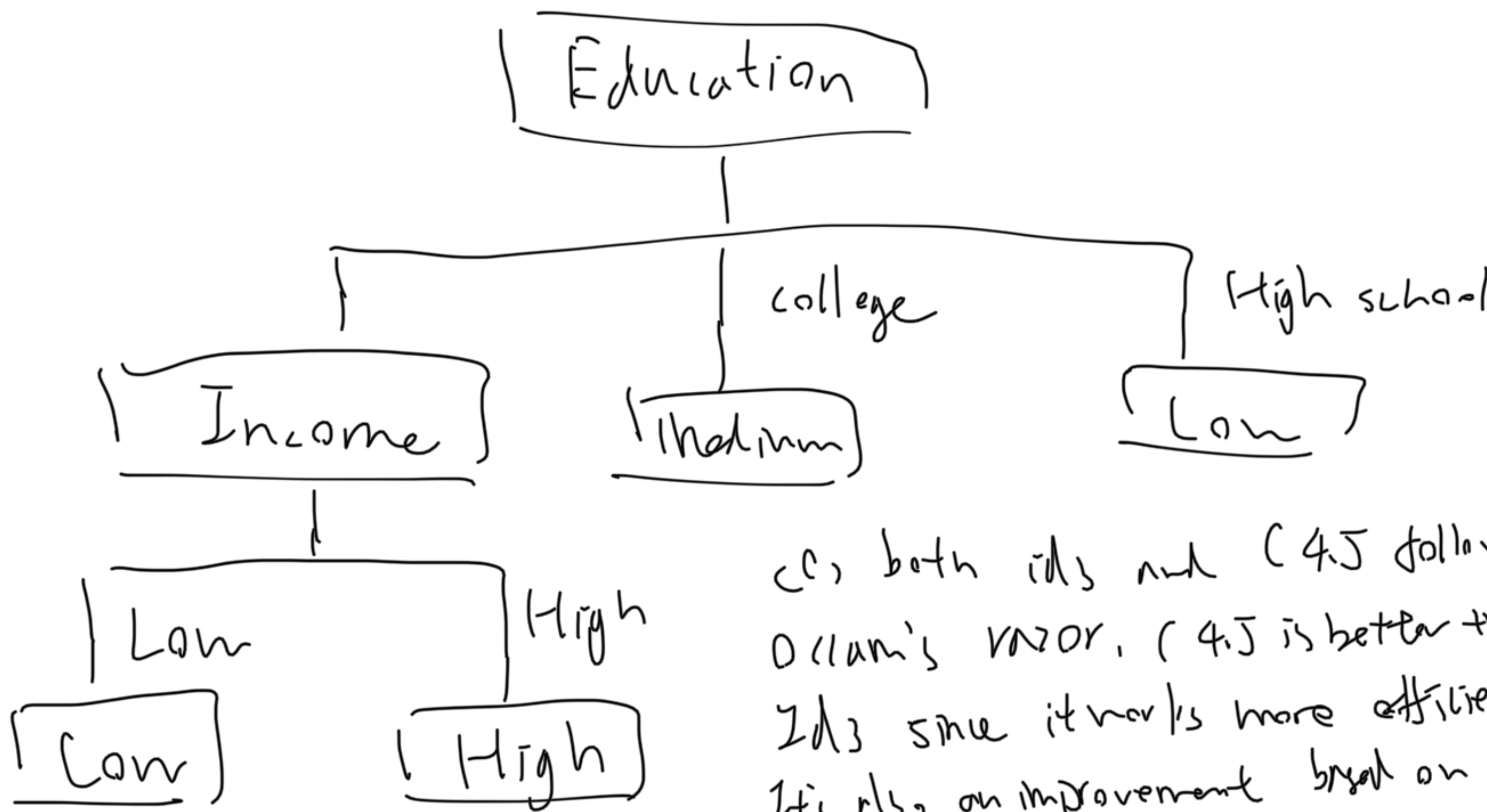
In this case, both marital status and Income could be chosen as the next node.

(iii) choice one - choose marital status as second node

$\boxed{\text{Education}}$

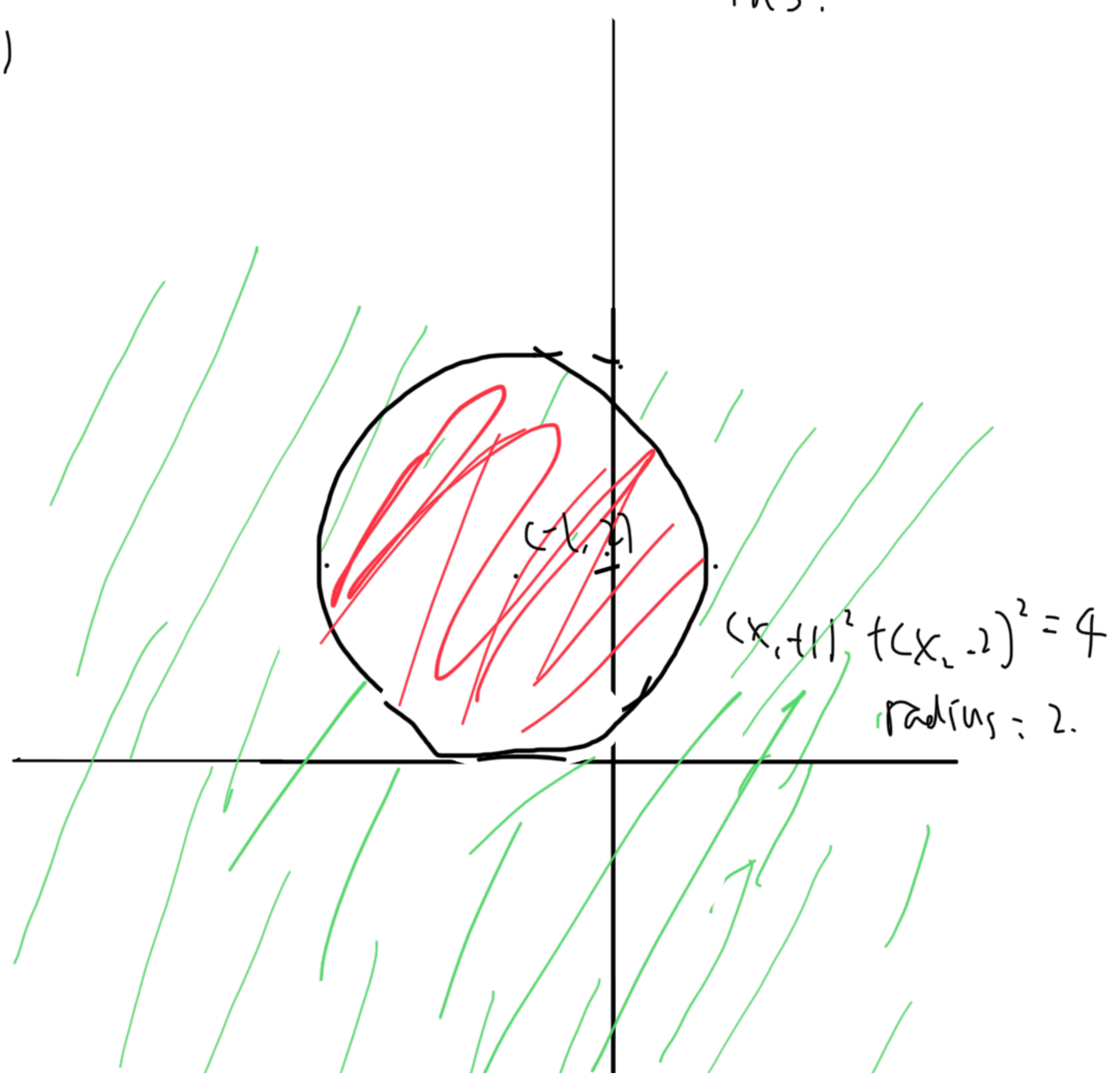


Choice two, Choose Income as the second node.



(c) both id3 and C4.5 follow Occam's razor. C4.5 is better than Id3 since it works more efficiently. It's also an improvement based on id3.

2. (a)



(b) the set of points for which $(1+X_1)^2+(2-X_2)^2>4$ is shaded by green

the set of points for which $(1+X_1)^2+(2-X_2)^2\le 4$ is shaded by red, with points on the circle $(1+X_1^2)^2+(2-X_2)^2=4$ included.

(c) observation $(0,0)$ will fall in green class
observation $(-1,1)$ will fall in red class
observation $(2,2)$ will fall in blue class
observation $(3,8)$ will fall in blue class

(d) $(1+X_1)^2+(2-X_2)^2=4$

$$X_1^2+2X_1+1+X_2^2-4X_2+4=4$$

$$1+2X_1+X_1^2-4X_2+X_2^2=0$$

As we can see that, through transformation, the decision boundary is in form

$$\beta_0+\beta_1X_1+\beta_2X_1^2+\beta_3X_2+\beta_4X_2^2=0$$

This is linear in terms of X_1, X_1^2, X_2, X_2^2, but not linear in terms of only X_1 and X_2.

3. (a)

b)

Graph (a) (b) (c) is linearly separable

(d) is linearly separable with one miss classification.

The change from 1-NN to SVM is illustrated through graph

(c) Higher order polynomial kernels such as quadratic kernel could be applied to figure (d) to make blue and red. points linearly separable.

4.(a) The absolute error Loss is

$L = |y - f(x)|$

and the epsilon insensitive loss function will become.

$L_\varepsilon(y, \hat{y}) = |y - \hat{y}|$ since $|y-\hat{y}|$ will always $\geq \varepsilon = 0$

$f(x) = w^T x + b$ $\quad |y - \hat{y}| = |y - f(x)|$

$\hat{y} = w^T x + b$

I would say, when $\varepsilon = 0$, the epsilon insensitive loss function is the same as the absolute error loss.

The ε's function is that, in epsilon insensitive loss function, all the errors $|y - \hat{y}|$ smaller than ε distance of the observed value will be treated as 0.

(b) $J(w) = \frac{1}{n} \sum_{i=1}^{n} L_\varepsilon(y, \hat{y}(x_i)) + \lambda \|w\|_2^2$

$= \frac{1}{n} \sum_{i=1}^{n} (|y - w^T x_i| - \varepsilon) + \lambda \|w\|_2^2$

add slack variable to the objective function.

$$= \frac{1}{n}\sum_{i=1}^{n}(|y - w^T x_i| - \varepsilon) + \lambda\|w\|_2 + \sum_{i=1} \xi_i$$

$$= \frac{1}{n}\sum_{i=1}^{n}(|y - w^T x_i| - \varepsilon + \xi_i) + \lambda\|w\|_2^2$$

since the constrint is

$$L_\varepsilon(y, \hat{y}(x_i)) = \begin{cases} 0 & \text{if } |y - \hat{y}(x_i)| < \varepsilon \\ |y - \hat{y}(x_i)| - \varepsilon & \text{, otherwise.} \end{cases}$$

I would like to add ξ_i to the constraint.

making it $L_\varepsilon(y, \hat{y}(x)) = 0$, $y - \hat{y}(x) \le \varepsilon + \xi_i$

i.e. $y - \hat{y}(x_i)$ is always smaller than $\varepsilon + \xi_i$.

making it always give 0 for $L_\varepsilon(y, \hat{y}(x_i))$

now $y - \hat{y}(x_i) \le \varepsilon + \xi_i$

$$-(y - \hat{y}(x_i)) \le \varepsilon + \xi_i$$

$$y - \hat{y}(x_i) \ge -\varepsilon - \xi_i$$

The optimization function becomes

$$J(w) = \frac{1}{n}\sum_{i=1}^{n}\xi_i + \lambda\|w\|_2^2$$

with constraint

$$y - \hat{y}(x_i) \le \varepsilon + \xi_i$$

$$y - \hat{y}(x_i) \ge -\varepsilon - \xi_i$$

and $\xi_i \ge 0$

This is an optimization problem that is differentiable and with linear constraints.