COMS W4721 Spring 2020 Homework 2: Linear Classifiers, Decision Trees

Shuyu Huang Sh3967

Instruction

Please prepare your write-up as a typeset PDF document (which can be generated using LaTex or Word). If you choose to hand-write certain portions of the assignment, make sure your handwriting is legible and the consistency in the format with other pages which are typeset (e.g. page size, page numbering). If we cannot read your handwriting, you may not receive credit for the question. This write-up contains all of your supporting materials which include plots, source code, and proofs for each of the parts in the assignment. Submit your assignment on Gradescope by clearly marking the pages for each part. On the first page of your write-up, please typeset: (1) your name and your UNI; (2) all of your collaborators whom you discussed the assignment with; (3) the parts of the assignment you had collaborated on. The solutions to the problems need to start from the second page. Please write up the solutions by yourself. The academic rules of conduct is found in the course syllabus.

Suggestions

If necessary, please define notations and explain reasoning behind the solutions as concisely as possible. Solutions without explanations when needed may receive no credit. Points can be deducted for solutions with unnecessarily long explanations for lack of clarity. Source code comment can be useful for explaining the logic behind your solutions. Please start early!

collabrator: Hanjun Li (hl3339) Q4

Problem 1: Linear Regression (20 points)

In our lecture, we discussed linear regression in which the data $\{(\mathbf{x}^{(i)} \in \mathbb{R}^d, y^{(i)})\}_{i=1}^N$ is generated such that $y^{(i)}$'s independent of others when conditioned on $\mathbf{x}^{(i)}$.

Let us assume that $p(y^{(i)} \mid \mathbf{x}^{(i)}; \mathbf{w}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2}{2\sigma^2}\right)$. In this problem we will explore the Bayesian formulation of linear regression called *maximum a posteriori estimate*:

$$
\begin{aligned}
\mathbf{w}_{\mathsf{MAP}} &= \arg\max_{\mathbf{w}} p(\mathbf{w} \mid \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N) \\
&= \arg\max_{\mathbf{w}} \frac{p(\{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N \mid \mathbf{w}) p(\mathbf{w})}{\int\limits_{\mathbf{w}'} p(\{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N \mid \mathbf{w}') p(\mathbf{w}') d\mathbf{w}'} \\
&= \arg\max_{\mathbf{w}} p(\{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N \mid \mathbf{w}) p(\mathbf{w}) \\
&= \arg\max_{\mathbf{w}} p(\{\mathbf{x}^{(i)}\}_{i=1}^N \mid \mathbf{w}) p(\{y^{(i)}\}_{i=1}^N \mid \{\mathbf{x}^{(i)}\}_{i=1}^N; \mathbf{w}) p(\mathbf{w}) \\
&= \arg\max_{\mathbf{w}} \ln p(\{\mathbf{x}^{(i)}\}_{i=1}^N \mid \mathbf{w}) + \ln p(\{y^{(i)}\}_{i=1}^N \mid \{\mathbf{x}^{(i)}\}_{i=1}^N; \mathbf{w}) + \ln p(\mathbf{w}) \\
&= \arg\max_{\mathbf{w}} \ln p(\{y^{(i)}\}_{i=1}^N \mid \{\mathbf{x}^{(i)}\}_{i=1}^N; \mathbf{w}) + \ln p(\mathbf{w}) \\
&= \arg\max_{\mathbf{w}} \left(\sum_{i=1}^N \ln p(y^{(i)} \mid \mathbf{x}^{(i)}; \mathbf{w}) \right) + \ln p(\mathbf{w})
\end{aligned}
$$

where the integral in the denominator can be dropped since it has no dependence on $\mathbf{w}$ and the natural log can be taken because it is a monotonic transform. For the last equation $\ln p(\{\mathbf{x}^{(i)}\}_{i=1}^N; \mathbf{w})$ is dropped because the data has no dependence on $\mathbf{w}$. Depending on the prior function on the weight $p(\mathbf{w})$, we can get several flavors of Bayesian linear regression.

(a) (10 points) Suppose each component of $\mathbf{w}$ is selected such that $w_i \sim N(0, \tau^2)$ i.i.d. What is $\mathbf{w}_{\mathsf{MAP}}$ in this case? Have we seen this form before?

(b) (10 points) Let us assume each component of $\mathbf{w}$ is selected such that $w^{(i)} \sim \text{Laplace}(0, b)$ i.i.d. What is $\mathbf{w}_{\mathsf{MAP}}$ in this case? Have we seen this form before?

1 (a) If each component $w_i \overset{iid}{\sim} N(0, \tau^2)$ in $\vec{w}$

Then, $P(w_i) = \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{w}{2\tau^2}}$

$P(\vec{w}) = \prod_{i=1}^{d} P(w_i) = (2\pi)^{-\frac{d}{2}} (\tau^2 I)^{-\frac{1}{2}} \exp\left[-\frac{\vec{w}^T\vec{w}}{2(\tau^2 I)}\right]$, let $\Sigma = \tau^2 I$

$\ln P(\vec{w}) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(\underset{\tau^2}{\Sigma}) - \frac{\vec{w}^T\vec{w}}{2\underset{\tau^2}{\Sigma}}$

By prompt,

$$P(y^{(i)} \mid \vec{x}^{(i)}; \vec{w}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y^{(i)} - \vec{w}^T\vec{x}^{(i)})^2}{2\sigma^2}\right)$$

$$\sum_{i=1}^{N} \ln P(y^{(i)} \mid \vec{x}^{(i)}; \vec{w}) = \sum_{i=1}^{N}\left(-\frac{1}{2}\ln(2\pi) - \ln\sigma - \frac{(y^{(i)} - \vec{w}^T\vec{x}^{(i)})^2}{2\sigma^2}\right)$$
$$= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \frac{\sum_{i=1}^{N}(y^{(i)} - \vec{w}^T\vec{x}^{(i)})^2}{2\sigma^2}$$

$$\vec{w}_{MAP} = \max_{\vec{w}}\left(\sum_{i=1}^{N}\ln P(y^{(i)} \mid \vec{x}^{(i)}; \vec{w}) + \ln P(\vec{w})\right)$$
$$= \max_{\vec{w}}\left(-\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \frac{\sum_{i=1}^{N}(y^{(i)} - \vec{w}^T\vec{x}^{(i)})^2}{2\sigma^2} - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(\Sigma) - \frac{\vec{w}^T\vec{w}}{2\Sigma}\right)$$
$$= \max_{\vec{w}} -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y^{(i)} - \vec{w}^T\vec{x}^{(i)})^2 - \frac{1}{2\Sigma}\vec{w}^T\vec{w}$$
$$= \min_{\vec{w}}\left(\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y^{(i)} - \vec{w}^T\vec{x}^{(i)})^2 + \frac{1}{2\Sigma}\vec{w}^T\vec{w}\right)\cdot\frac{2\sigma^2}{N}$$
$$= \min_{\vec{w}}\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \vec{w}^T\vec{x}^{(i)})^2 + \frac{\sigma^2}{N\Sigma}\vec{w}^T\vec{w}$$
where $\frac{\sigma^2}{N\Sigma} = \frac{\sigma^2}{N\tau^2}$

Note that $\hat{w}_{ridge} = \min_{\vec{w}}\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \vec{w}^T\vec{x}^{(i)}) + \underbrace{\lambda}_{\geq 0}\vec{w}^T\vec{w}$

Therefore it has same form of $\vec{w}_{ridge}$ in ridge regression where $\lambda = \frac{\sigma^2}{N\tau^2}$

1 (b) If $w_i \overset{iid}{\sim}$ Laplace $(0, b)$

Then $P(w_i) = \frac{1}{2b}\exp\left[-\frac{|w_i|}{b}\right]$

$P(\vec{w}) = \prod_{i=1}^{d}\frac{1}{2b}\exp\left[-\frac{|w_i|}{b}\right] = (2b)^{-d}\exp\left[-\frac{1}{b}\sum_{i=1}^{d}|w_i|\right]$

$\ln P(\vec{w}) = -d\ln(2b) - \frac{1}{b}\sum_{i=1}^{d}|w_i|$

By prompt,

$$P(y^{(i)} \mid \vec{x}^{(i)}; \vec{w}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y^{(i)} - \vec{w}^T\vec{x}^{(i)})^2}{2\sigma^2}\right)$$

$$\sum_{i=1}^{N} \ln P(y^{(i)} \mid \vec{x}^{(i)}; \vec{w}) = \sum_{i=1}^{N}\left(-\frac{1}{2}\ln(2\pi) - \ln\sigma - \frac{(y^{(i)} - \vec{w}^T\vec{x}^{(i)})^2}{2\sigma^2}\right)$$
$$= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \frac{\sum_{i=1}^{N}(y^{(i)} - \vec{w}^T\vec{x}^{(i)})^2}{2\sigma^2}$$

$$\vec{w}_{MAP} = \max_{\vec{w}}\left(\sum_{i=1}^{N}\ln P(y^{(i)} \mid \vec{x}^{(i)}; \vec{w}) + \ln P(\vec{w})\right)$$
$$= \max_{\vec{w}}\left(-\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \frac{\sum_{i=1}^{N}(y^{(i)} - \vec{w}^T\vec{x}^{(i)})^2}{2\sigma^2} - d\ln(2b) - \frac{1}{b}\sum_{i=1}^{d}|w_i|\right)$$
$$= \max_{\vec{w}} -\frac{\sum_{i=1}^{N}(y^{(i)} - \vec{w}^T\vec{x}^{(i)})^2}{2\sigma^2} - \frac{1}{b}\underbrace{\sum_{i=1}^{d}|w_i|}$$
$$= \min_{\vec{w}}\left(\frac{\sum_{i=1}^{N}(y^{(i)} - \vec{w}^T\vec{x}^{(i)})^2}{2\sigma^2} + \frac{1}{b}\sum_{i=1}^{d}|w_i|\right)\frac{2\sigma^2}{N}$$
, where $\sum_{i=1}^{d}|w_i| = \|\vec{w}\|_1$,
$$= \min_{\vec{w}}\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \vec{w}^T\vec{x}^{(i)})^2 + \frac{2\sigma^2}{Nb}\|\vec{w}\|_1$$

Note that $\hat{w}_{lasso} = \min_{\vec{w}}\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \vec{w}^T\vec{x}^{(i)})^2 + \underbrace{\lambda}_{\geq 0}\|\hat{w}\|_1$

Therefore it has same form of $\hat{w}_{lasso}$ in LASSO regression where $\lambda = \frac{2\sigma^2}{Nb}$

Problem 2: Gaussian Discriminant Analysis (15 points)

(a) (5 points) Prove that a Gaussian discriminant analysis in $\mathbb{R}^d$ induces a quadratic decision boundary. You can assume a binary classification task for the purpose of this task. What properties of the parameters will ensure a linear decision boundary?

(b) (10 points) Let's take the Gaussian discriminant analysis in $\mathbb{R}^d$ from part (a). What properties of the parameters can make the classifier a Gaussian Naive Bayes Classifier? Give a short mathematical proof to validate your answer.

2(a) Guassian Discriminant Analysis:

Data:

$\{(\vec{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$

$\vec{x}^{(i)} \in \mathcal{X} \in \mathbb{R}^d$, $y^{(i)} \in \mathcal{Y} (:= \{0,1\})$

Assumption:

$$P(\vec{x} \mid y; \mu_y, \Sigma_y) = (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left[\frac{-(\vec{x}-\mu_y)^T(\Sigma_y)^{-1}(\vec{x}-\mu_y)}{2}\right]$$

In GDA,

$$\hat{f}(\vec{x}) = \max_{y \in \{0,1\}} P(y \mid \vec{x}) = \max_{y \in \{0,1\}} \pi_y P(\vec{x} \mid y; \mu_y, \Sigma_y)$$

let $d_\Sigma(\vec{x}, \mu) = (\vec{x}-\mu)^T(\Sigma)^{-1}(\vec{x}-\mu)$

Then

$$\hat{f}(\vec{x}) = \mathbb{1}[P(y=1 \mid \vec{x}) > P(y=0 \mid \vec{x})$$

$$= \mathbb{1}\left[\ln \frac{P(y=1 \mid \vec{x})}{P(y=0 \mid \vec{x})} > 0\right]$$

$$= \mathbb{1}\left[\ln \frac{P(y=1)}{1-P(y=1)} + \ln \frac{P(\vec{x} \mid y=1)}{P(\vec{x} \mid y=0)} > 0\right]$$

let $\pi_1 = P(y=1)$

$$= \mathbb{1}\Big[\underbrace{\ln \frac{\pi_1}{1-\pi_1}}_{①} - \frac{1}{2}\underbrace{\ln \frac{|\Sigma_1|}{|\Sigma_2|}}_{②} - \frac{1}{2}\underbrace{(d_{\Sigma_1}(\vec{x}, \mu_1) - d_{\Sigma_0}(\vec{x}, \mu_2))}_{③} > 0\Big]$$

① is constant since $\pi_1 = P(y=1)$ is independent of $\vec{x}$

② is constant since Σ_1, Σ_2 are covariance matries

③ is scale differences.

Since $d_\Sigma(\vec{x}, \mu) = (\vec{x}-\mu)^T(\Sigma)^{-1}(\vec{x}-\mu)$, where Σ is covariance matrix, μ is mean it is a quadratic term in terms of $\vec{x}$.

Therefore, Guassian Discriminant Analysis has a quadratic boundary.

If we want a linear decision boundary, we need $\boxed{\Sigma_0 = \Sigma_1}$

Suppose $\Sigma_0 = \Sigma_1$

$$\hat{f}(\vec{x}) = \mathbb{1}\left[\ln \frac{\pi_1}{1-\pi_1} - \cancel{\frac{1}{2}\ln \frac{|\Sigma_1|}{|\Sigma_2|}} - \frac{1}{2}(d_{\Sigma_1}(\vec{x}, \mu_1) - d_{\Sigma_0}(\vec{x}, \mu_2)) > 0\right]$$

$$d_{\Sigma_1}(\vec{x}, \mu_1) - d_{\Sigma_2}(\vec{x}, \mu_2) = (\vec{x}-\mu_1)^T(\Sigma_1)^{-1}(\vec{x}-\mu_1) - (\vec{x}-\mu_0)^T(\Sigma_1)^{-1}(\vec{x}-\mu_0)$$

$$= \vec{x}^T \underbrace{(\Sigma_1)^{-1}(\mu_1-\mu_2)}_{\text{constant}} - \frac{1}{2}\mu_1^T(\Sigma_1)^{-1}\mu_1 + \frac{1}{2}\mu_2^T(\Sigma_1)^{-1}\mu_0 \quad \text{is linear in } \vec{x}$$

2 (b) Guassian Naive Bayes Classifier.

$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$, where $x_k \in \mathbb{R}$

$$\hat{f}_{GNBC}(x) = \max_{y \in Y} \prod_{k=1}^{d} P(x_k | y; \mu_k, \sigma_k^2) \cdot Pr(y)$$

when Σ_1 and Σ_2 are diagonal matrices, Guassian Naive Bayes Classifier is a special case of GDA

Proof.

GDA has

$$\hat{f}_{GDA}(\vec{x}) = \max_{y \in \{0,1\}} P(y|\vec{x}) = \max_{y \in \{0,1\}} \pi_y \, p(\vec{x}|y; \mu_y, \Sigma_y)$$

where $p(\vec{x}|y; \mu_y, \Sigma_y) = (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left[\frac{-(\vec{x}-\mu_y)^T (\Sigma_y)^{-1} (\vec{x}-\mu_y)}{2}\right]$

by 2(a)

$$\hat{f}_{GDA}(\vec{x}) = \mathbb{1}\left[\ln \frac{P(y=1)}{1-P(y=1)} - \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|} - \frac{1}{2}\left(d_{\Sigma_1}(\vec{x}, \mu_1) - d_{\Sigma_0}(\vec{x}, \mu_2)\right) > 0\right]$$

$$= \mathbb{1}\left[\ln \frac{P(y=1)}{1-P(y=1)} - \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|} - \frac{1}{2}\left[(\vec{x}-\mu_1)^T (\Sigma_1)^{-1} (\vec{x}-\mu_1) - (\vec{x}-\mu_2)^T (\Sigma_2)^{-1} (\vec{x}-\mu_2)\right] > 0\right]$$

Suppose Σ_1, Σ_2 are diagonal matrix

$$\Sigma_1 = \begin{bmatrix} \sigma_{11}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{12}^2 & & \vdots \\ 0 & 0 & \ddots & \\ \vdots & & & \\ 0 & \cdots & & \sigma_{1d}^2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} \sigma_{21}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{22}^2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \sigma_{2d}^2 \end{bmatrix} \Rightarrow (\Sigma_1)^{-1} = \begin{bmatrix} \sigma_{11}^{-2} & 0 & \cdots & 0 \\ 0 & \sigma_{12}^{-2} & & \vdots \\ 0 & 0 & \ddots & \\ \vdots & & & \\ 0 & \cdots & & \sigma_{1d}^{-2} \end{bmatrix} \quad (\Sigma_2)^{-1} = \begin{bmatrix} \sigma_{21}^{-2} & 0 & \cdots & 0 \\ 0 & \sigma_{22}^{-2} & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \sigma_{2d}^{-2} \end{bmatrix}$$

(Σ_1, Σ_2: d rows; $(\Sigma_1)^{-1}$: d rows)

$$|\Sigma_1| = \prod_{i=1}^{d} \sigma_{1i}^2 \qquad |\Sigma_2| = \prod_{i=1}^{d} \sigma_{2i}^2$$

Then

$$(\vec{x}-\mu_1)^T (\Sigma_1)^{-1} (\vec{x}-\mu_1)$$

$$= \begin{bmatrix} (x_1-\mu_{11}) & (x_2-\mu_{12}) & \cdots & (x_d-\mu_{1d}) \end{bmatrix} \begin{bmatrix} \sigma_{11}^{-2} & 0 & \cdots & 0 \\ 0 & \sigma_{12}^{-2} & & \vdots \\ 0 & 0 & \ddots & \\ \vdots & & & \\ 0 & \cdots & & \sigma_{1d}^{-2} \end{bmatrix} \begin{bmatrix} x_1-\mu_{11} \\ x_2-\mu_{12} \\ \vdots \\ x_d-\mu_{1d} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{x_1-\mu_{11}}{\sigma_{11}^2} & \frac{x_2-\mu_{12}}{\sigma_{12}^2} & \cdots & \frac{x_d-\mu_{1d}}{\sigma_{1d}^2} \end{bmatrix} \begin{bmatrix} x_1-\mu_{11} \\ \\ \\ x_d-\mu_{1d} \end{bmatrix}$$

$$= \left[\sum_{i=1}^{d} \left(\frac{x_i-\mu_{1i}}{\sigma_{1i}}\right)^2\right]$$

Similarly,

$$(\vec{x}-\mu_2)^T (\Sigma_2)^{-1} (\vec{x}-\mu_2) = \left[\sum_{i=1}^{d} \left(\frac{x_i-\mu_{2i}}{\sigma_{2i}}\right)^2\right]$$

SO $$\hat{f}_{GDA}(\vec{x}) = \mathbb{1}\left[\ln \frac{P(y=1)}{1-P(y=1)} - \frac{1}{2}\ln\prod_{i=1}^{d}\frac{\sigma_{1i}^2}{\sigma_{2i}^2} - \frac{1}{2}\left[\sum_{i=1}^{d}\left(\left(\frac{x_i-\mu_{1i}}{\sigma_{1i}}\right)^2 - \left(\frac{x_i-\mu_{2i}}{\sigma_{2i}}\right)^2\right) > 0\right]\right]$$

$$= \mathbb{1}\left[\ln \frac{P(y=1)}{1-P(y=1)} - \sum_{i=1}^{d}\ln\frac{\sigma_{1i}}{\sigma_{2i}} - \frac{1}{2}\left[\sum_{i=1}^{d}\left(\left(\frac{x_i-\mu_{1i}}{\sigma_{1i}}\right)^2 - \left(\frac{x_i-\mu_{2i}}{\sigma_{2i}}\right)^2\right) > 0\right]\right]$$

– continue in next page –

$$\hat{f}_{GNBC}(x) = \max_{y \in y} \prod_{k=1}^{d} \underbrace{P(X_k | y; \mu_k, \sigma_k^2)}_{\text{univariate Guassian}} \cdot Pr(y)$$

$$= \mathbb{1}\left[P(y) \left\{ \prod_{k=1}^{d} P(X_k | y=1; \mu_{1k}, \sigma_{1k}^2) > P(X_k | y=0; \mu_{2k}, \sigma_{2k}^2) \right\} \right]$$

$$= \mathbb{1}\left[\ln \frac{P(y)}{1-P(y)} + \ln \frac{\prod_{k=1}^{d} P(X_k | y=1; \mu_{1k}, \sigma_{1k}^2)}{\prod_{k=1}^{d} P(X_k | y=0; \mu_{2k}, \sigma_{2k}^2)} > 0 \right]$$

$$= \mathbb{1}\left[\ln \frac{P(y)}{1-P(y)} + \left\{ \sum_{k=1}^{d} \ln P(X_k | y=1; \mu_{1k}, \sigma_{1k}^2) - \sum_{k=1}^{d} \ln P(X_k | y=0; \mu_{2k}, \sigma_{2k}^2) > 0 \right\} \right]$$

Note: $P(\vec{x} | y; \mu_y, \Sigma_y) = (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left[\frac{-(\vec{x}-\mu_y)^T (\Sigma_y)^{-1} (\vec{x}-\mu_y)}{2} \right]$

$$= \mathbb{1}\left[\ln \frac{P(y)}{1-P(y)} + \sum_{k=1}^{d} \ln \frac{\sigma_{1k}}{\sigma_{2k}} - \frac{1}{2} \left[\sum_{k=1}^{d} \left(\left(\frac{x_k - \mu_{1k}}{\sigma_{1k}} \right)^2 - \left(\frac{x_k - \mu_{1k}}{\sigma_{2k}} \right)^2 \right) > 0 \right] \right]$$

$$= \hat{f}_{GDA}(\vec{x})$$

Hence, we proved that when Σ_1, Σ_2 are diagonal matrices, GDA can be turned into Guassian Naive Bayes classifier.

Problem 3: Logistic Regression (30 points)

(a) (15 points) Logistic Regression can be used to solve the binary classification task as depicted in the figure. A simple logistic regression model is given below:

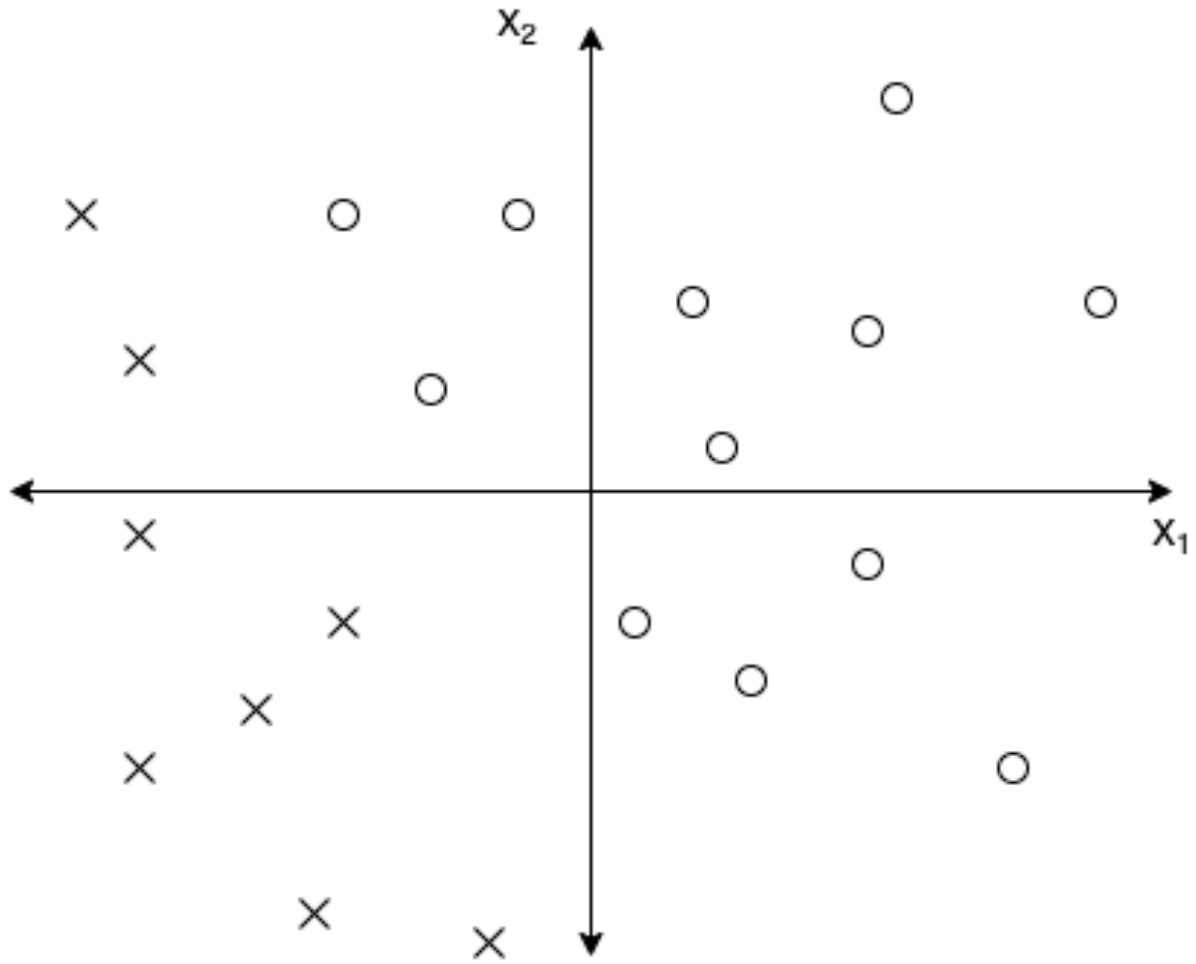


Figure 1: Binary classification dataset

$$P(y = class1 \mid \vec{x}, \vec{w}) = \frac{1}{1 + exp(-w_0 - w_1 x_1 - w_2 x_2)}$$

The given data is linearly separable at this point and thus a logistic regression model can be fit to separate the data with zero training error.
Consider a regularized logistic regression model where out optimization (maximization in this case) function becomes:

$$\sum_{i=1}^{N} log(P(y^{(i)} \mid x^{(i)}; w_0, w_1, w_2)) - \lambda \cdot w_j^2$$

for a very large λ. The regularization penalties used penalize one parameter at a time, ie. either one of $\{w_0, w_1, w_2\}$ are penalized at a given time. How does the training error change with regularization of each parameter? Provide a brief mathematical justification for each of your answers.

(b) (15 points) You are given a dataset *logistic_regression.csv* which has two features x_1, x_2 and the corresponding *class* label. Please write a small program in a language of your choice to optimize the logistic regression function to

- Fit the data in the csv file without regularization
- Regularize the squared weight of w_1 associated with the feature x_1.
  For each value of $\lambda \in [10^0, 10^1, 10^2, 10^3, 10^4, 10^5]$: plot the decision boundary along with the points in the dataset. You should have 6 plots.
- Regularize the squared weight of w_2 associated with the feature x_2.
  For each value of $\lambda \in [10^3, 10^4, 10^5, 10^6, 10^7, 10^8]$: plot the decision boundary along with the points in the dataset. You should have 6 plots.

Please attach all plots for all of the three subparts with your analysis. You may use scipy.optimize.minimize for implementing your code. You do not need to include the code. You may find the Python notebook for problem 3 useful for the starter code (please see section with "Modify Me"'s).

3(a)

Since λ is very large, $\sum_{i=1}^{N} \log\left(P(y^{(i)} \mid x^{(i)}; w_0, w_1, w_2)\right) - \lambda \cdot w_j^2 \xrightarrow[\lambda \text{ large}]{P} -\lambda w_j^2$

Regularize the loglikelihood:

$$\hat{w} = \max_{\vec{w}} \sum_{i=1}^{N} \log\left(P(y^{(i)} \mid x^{(i)}; w_0, w_1, w_2)\right) - \lambda \cdot w_j^2$$

$$\xrightarrow[\lambda \text{ very large}]{P} \max_{\vec{w}} -\lambda w_j^2$$

when $w_j = 0$, we obtain $\hat{w}$.

If we penalize / regularize on

① w_0 (intercept): as λ very large, w_0 will be close to 0, then the decision boundary will be a straight line pass through origin. Since "o" span on three quadrant, there doesn't exist a decision boundary splitting the two classes in the training set with no error. Therefore the training error will increase if we regularize on w_0.

② w_1: as λ very large, w_1 will be close to 0, then the decision boundary will be a straight horizontal line. Since two classes in the training set span over both $x_2 > 0$ and $x_2 < 0$, the training error would be high (largest among the three regularization choices. if we split them horizontally.

③ w_2 as λ very large, w_2 will be close to 0, then the decision boundary will be a straight vertical line. According to plot, dividing two classes with a verticle line will always introduce training error. However, we can observe a verticle gap between two gaps. It will give minimal training error among the three regularization choices.

In [9]:
```
import numpy
from matplotlib import colors
import matplotlib.pyplot as plt
from scipy import special
from scipy.optimize import minimize
import pandas as pd
```

In [10]:
```
# Register the color map to be used for plotting.
cmap = colors.LinearSegmentedColormap(
    'red_blue_classes',
    {'red': [(0, 1, 1), (1, 0.7, 0.7)],
     'green': [(0, 0.7, 0.7), (1, 0.7, 0.7)],
     'blue': [(0, 0.7, 0.7), (1, 1, 1)]})
plt.cm.register_cmap(cmap=cmap)
```

In [11]:
```
# Read the data from the problem 3 file.
from numpy import genfromtxt
my_data = genfromtxt('logistic_regression.csv', delimiter=',', skip_he
ader=1)
print(my_data.shape)
X = my_data[:, [0, 1]]
y = my_data[:, 2]

# Divide the data into two classes for plotting.
X0 = my_data[numpy.where(my_data[:, 2] == 0)]
X1 = my_data[numpy.where(my_data[:, 2] == 1)]
```

```
(2000, 3)
```

In [12]:
```
# Implements the natural log of the logistic function on $w^T x$.
def log_logistic_prob(iterate, x):
    t = numpy.dot(iterate[1:], x) + iterate[0]
    if t < -33:
        return t
    elif t < -18:
        return t - numpy.exp(t)
    elif t < 37:
        return -numpy.log1p(numpy.exp(- t ))
    else:
        return -numpy.exp(-t)

# Implements the logistic function on $w^T x$.
def logistic_prob(iterate, x):
    t = numpy.dot(iterate[1:], x) + iterate[0]
    if t < -33.3:
        return numpy.exp(t)
    elif t <= -18:
        return numpy.exp(t - numpy.exp(t))
    elif t <= 37:
        return numpy.exp(-numpy.log1p(numpy.exp(-t)))
    else:
        return numpy.exp(-numpy.exp(-t))
```

In [13]:
```
# Evaluates the logistic function on a set of grid points.
def logistic_prob_grid(iterate, grids):
    return numpy.array([logistic_prob(iterate, x) for x in grids ])

# Takes the logistic probability and thresholds to outputs a classific
ation label.
def logistic_pred(iterate, x):
    return 1.0 if logistic_prob(iterate, x) > 0.5 else 0.0
```

In [23]:
```
# Implements the negative logistic regression objective. Modify me!
def negative_log_likelihood(iterate, X, y, regularization):

    # The accumulated objective value.
    obj_val = 0.0


    # Loop over each (x, y) pair.
    for i, (x_vec, y) in enumerate(zip(X, y)):

        # Dot product $w^T x$.
        predict = iterate[0] + numpy.dot(iterate[1:], x_vec)

        # Accumulate the objective value contribution from this (x, y)
pair.
        obj_val += (- (1 - y) * predict + log_logistic_prob(iterate, x
_vec) )

    # Subtract the regularization parameter.
    #return - obj_val + regularization * numpy.dot(iterate[1:], iterat
e[1:])
    return -obj_val + numpy.dot(regularization*iterate[1:], iterate[1:
])
```

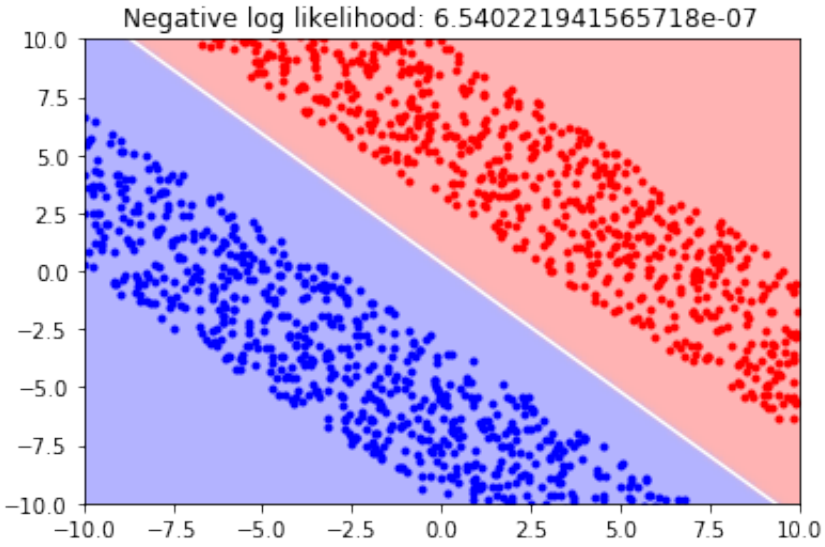
In [24]:
```
# Implements the logistic regression gradient. Modify me!
def gradient_negative_log_likelihood(iterate, X, y, regularization):
    gradient = numpy.zeros(3)

    # Loop over each (x, y) pair.
    for i, (x_vec, y) in enumerate(zip(X, y)):

        # Dot product $w^T x$.
        predict = iterate[0] + numpy.dot(iterate[1:], x_vec)

        if predict > 0.0:
            factor = ((y - 1) + y * numpy.exp(-predict)) / ( 1 + numpy
.exp(-predict) )
        else:
            factor = ((y - 1) * numpy.exp(predict) + y ) / ( 1 + numpy
.exp(predict) )
        gradient[0] -= factor
        gradient[1:] -= factor * x_vec

    # Regularize gradient.
    gradient[1:] += 2 * regularization * iterate[1:]

    return gradient
```

In [25]:
```
# Plots the data with the decision boundary.
def plot_data(X, y, iterate, regularization):
    y_pred = [logistic_pred(iterate, x) for x in X]
    tp = (y == y_pred)  # True Positive
    tp0, tp1 = tp[y == 0], tp[y == 1]
    X0, X1 = X[y == 0], X[y == 1]
    X0_tp, X0_fp = X0[tp0], X0[~tp0]
    X1_tp, X1_fp = X1[tp1], X1[~tp1]

    # class 0: dots
    plt.scatter(X0_tp[:, 0], X0_tp[:, 1], marker='.', color='red')
    plt.scatter(X0_fp[:, 0], X0_fp[:, 1], marker='x',
                s=20, color='#990000')  # dark red

    # class 1: dots
    plt.scatter(X1_tp[:, 0], X1_tp[:, 1], marker='.', color='blue')
    plt.scatter(X1_fp[:, 0], X1_fp[:, 1], marker='x',
                s=20, color='#000099')  # dark blue

    # class 0 and 1 : areas
    nx, ny = 200, 200
    x_min, x_max = (-10, 10)
    y_min, y_max = (-10, 10)
    plt.xlim(-10, 10)
    plt.ylim(-10, 10)
    xx, yy = numpy.meshgrid(numpy.linspace(x_min, x_max, nx),
                            numpy.linspace(y_min, y_max, ny))

    Z = logistic_prob_grid(iterate, numpy.c_[xx.ravel(), yy.ravel()])
    Z = Z.reshape(xx.shape)
    neg_log_likelihood = negative_log_likelihood(iterate, X, y, regula
rization)
    plt.title(
        'Negative log likelihood: ' + str(neg_log_likelihood))
    plt.pcolormesh(xx, yy, Z, cmap='red_blue_classes',
                   norm=colors.Normalize(0., 1.), zorder=0)

    # Plot my linear decision boundary here!
    linex = numpy.linspace(x_min, x_max, nx)
    liney = -iterate[0]/iterate[2]-iterate[1]/iterate[2]*linex
    plt.plot(linex, liney, c = 'white')



    plt.show()
```

In [26]:
```
# Plot for each different regularization value by calling your optimiz
ation routine for
# different values.
```

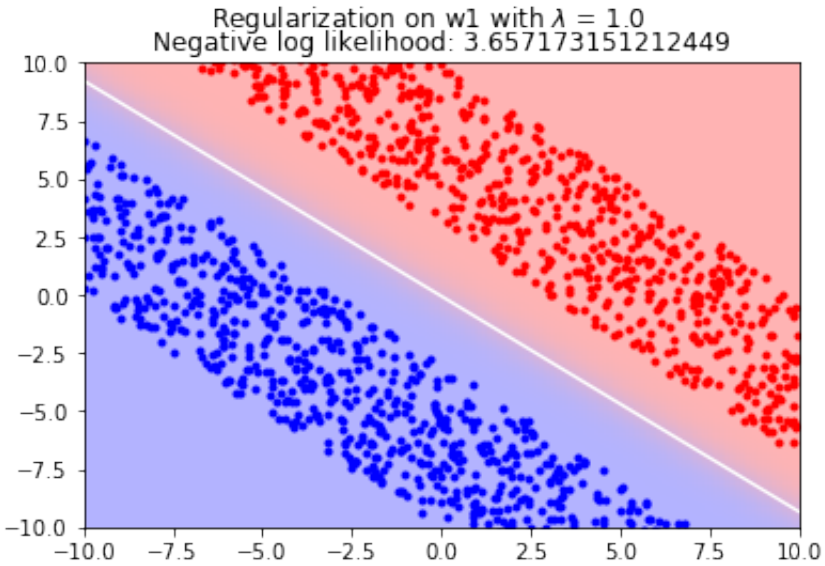
No Regularization

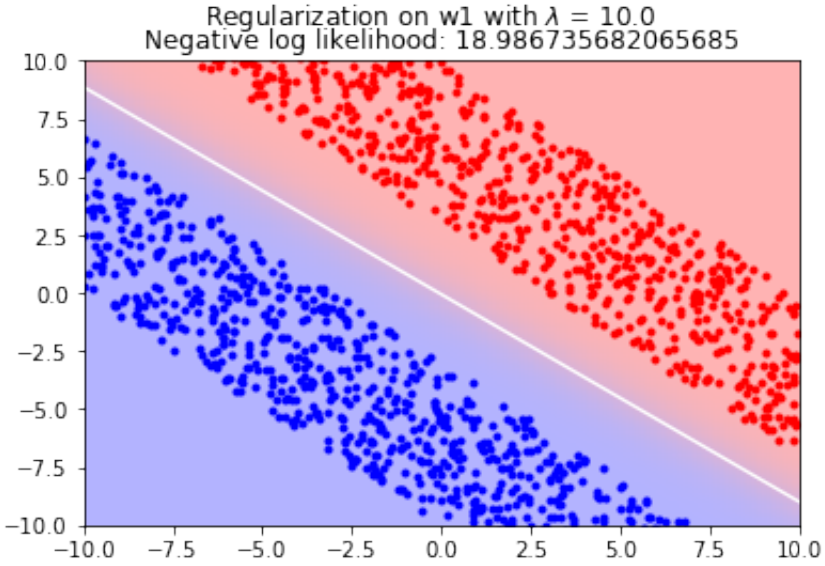
In [32]:
```
res0 = minimize(
    negative_log_likelihood,
    [0,0,0], method = 'BFGS',
    jac = gradient_negative_log_likelihood,
    args=(X,y,numpy.array([0,0]),))
plot_data(X,y,res0.x,0)#lambda=0 no regularization
```

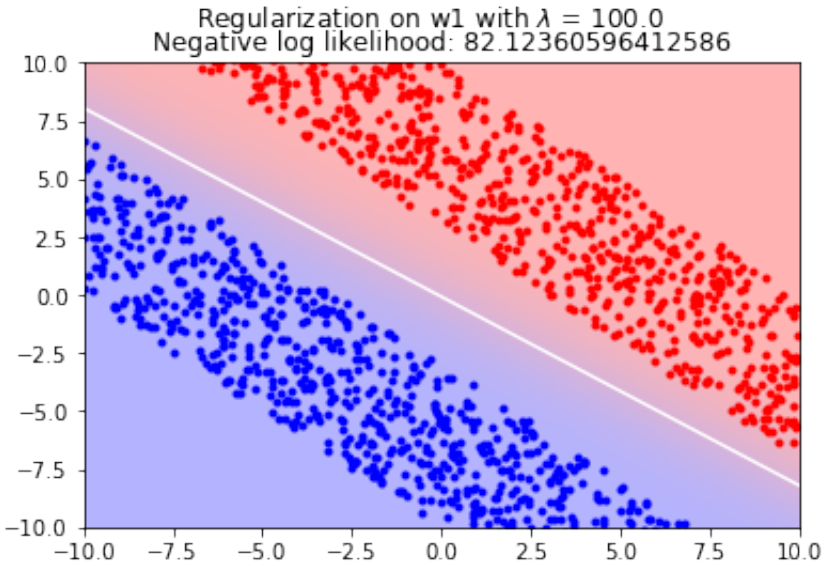


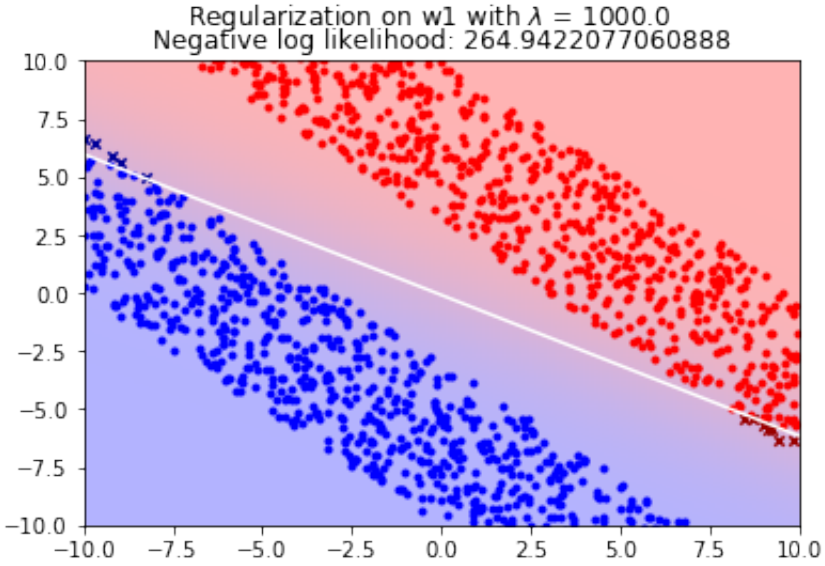
regularize on w_1 with $\lambda = [10^0, 10^1, 10^2, 10^3, 10^4, 10^5]$

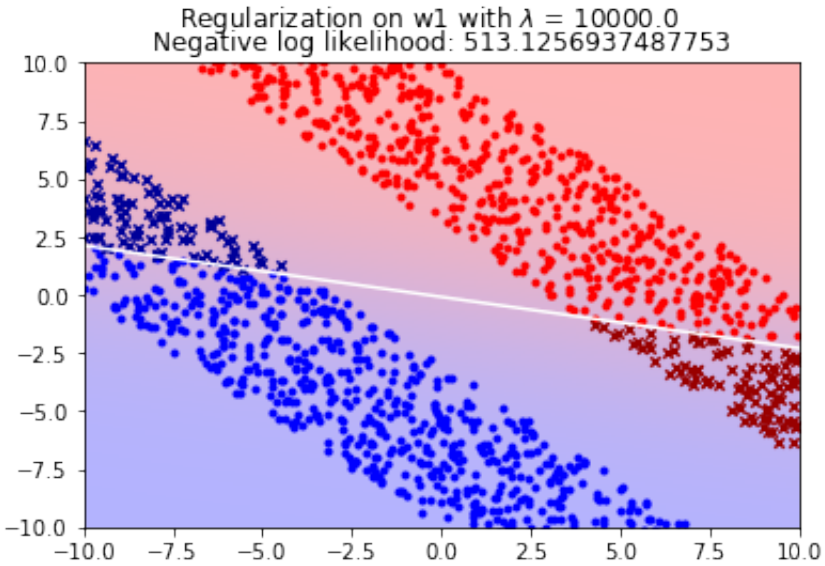
In [51]:
```
lambda_grid_w1 = numpy.logspace(0,5,6)
for i in range(6):
    regularization = lambda_grid_w1[i]
    res1 = minimize(negative_log_likelihood,[0,0,0],
                    method = 'BFGS',
                    jac = gradient_negative_log_likelihood,
                    args=(X,y,numpy.array([regularization,0])))
    plt.suptitle('Regularization on w1 with $\lambda$ = '+ str(regular
ization))
    plot_data(X,y,res1.x,numpy.array([regularization,0]))
```

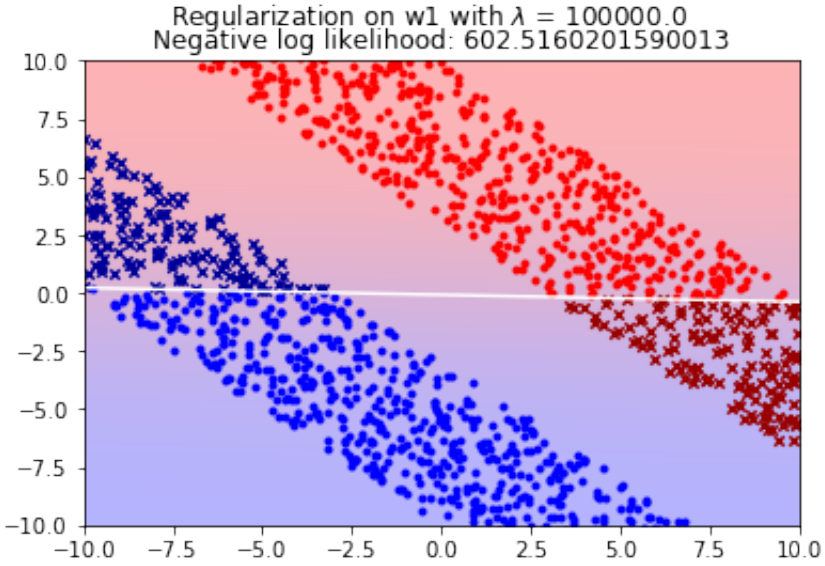

Regularization on w1 with λ = 1.0
Negative log likelihood: 3.657173151212449
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0 −7.5 −5.0 −2.5 0.0 2.5 5.0 7.5 10.0


Regularization on w1 with λ = 10.0
Negative log likelihood: 18.986735682065685
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0 −7.5 −5.0 −2.5 0.0 2.5 5.0 7.5 10.0


Regularization on w1 with λ = 100.0
Negative log likelihood: 82.12360596412586
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0 −7.5 −5.0 −2.5 0.0 2.5 5.0 7.5 10.0

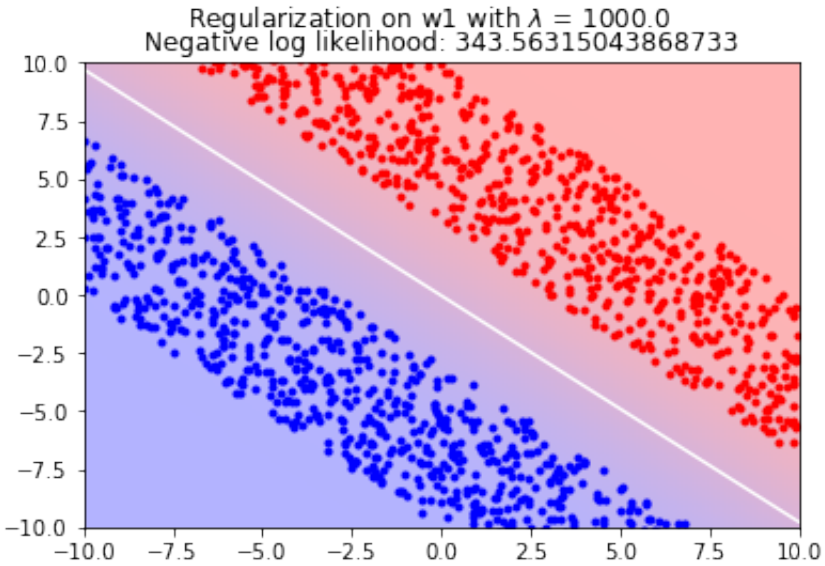

Regularization on w1 with λ = 1000.0
Negative log likelihood: 264.9422077060888
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0 −7.5 −5.0 −2.5 0.0 2.5 5.0 7.5 10.0

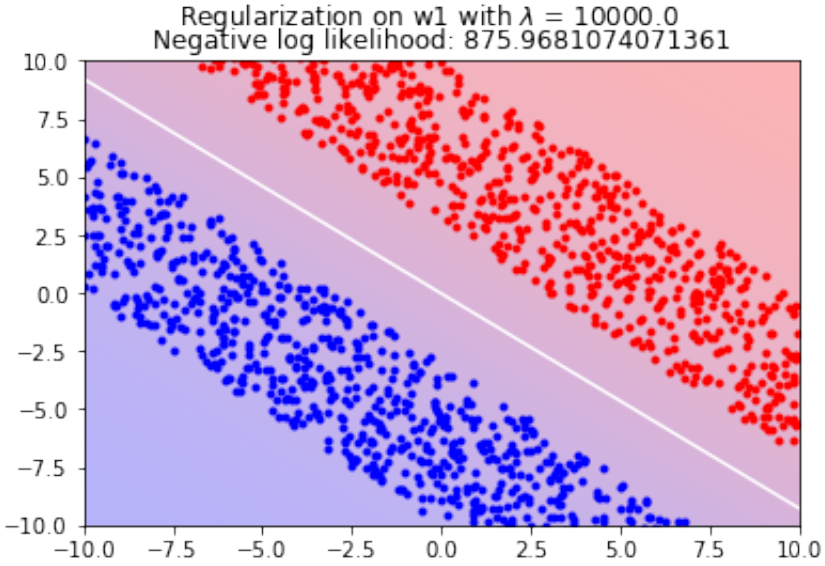


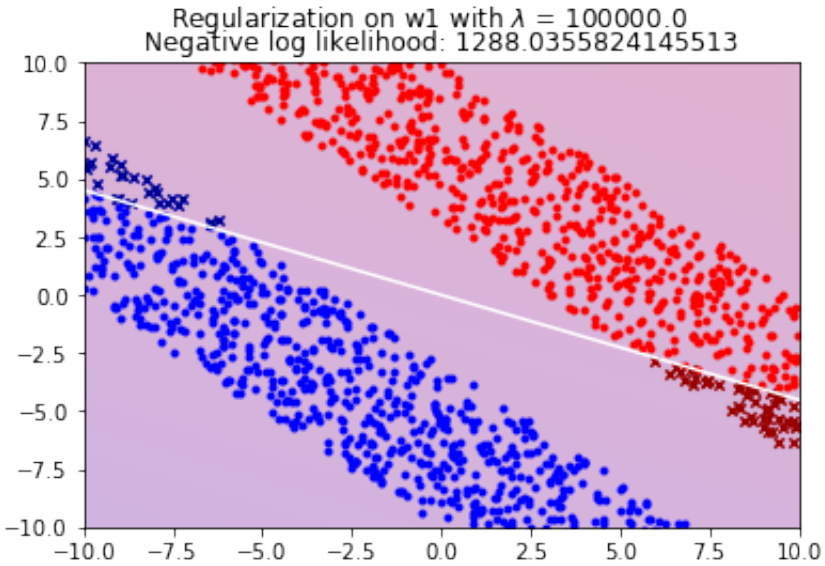


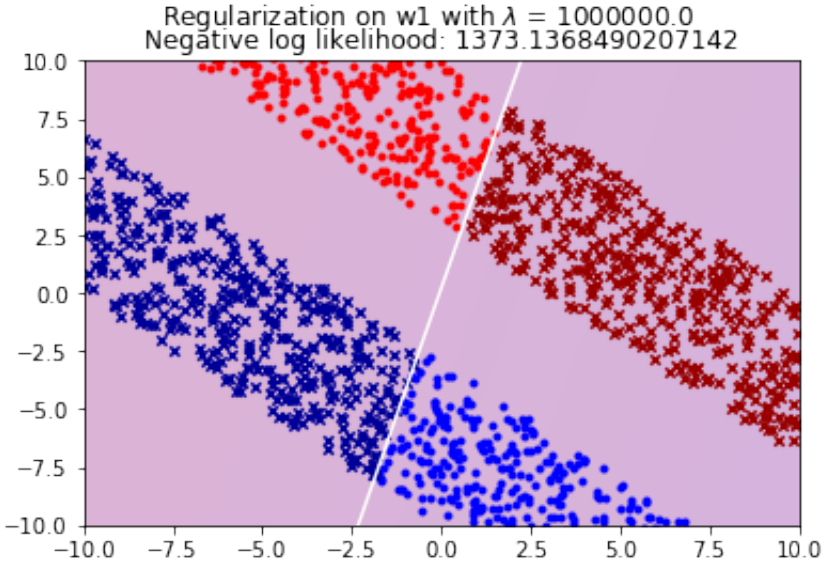
regularize on w_2 with $\lambda = [10^3, 10^4, 10^5, 10^6, 10^7, 10^8]$

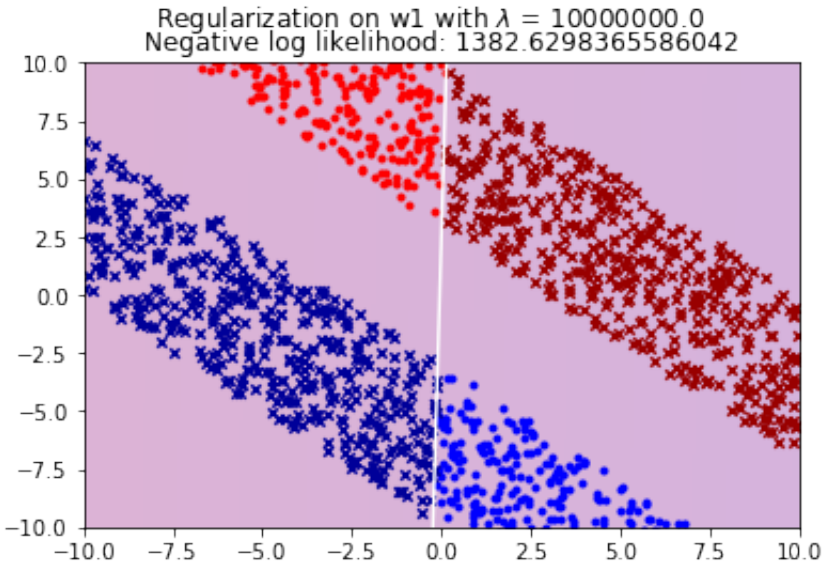
In [52]:
```
lambda_grid_w2 = numpy.logspace(3, 8, 6)
for i in range(6):
    regularization = lambda_grid_w2[i]
    res2 = minimize(negative_log_likelihood,[0,0,0],
                    method = 'BFGS',
                    jac = gradient_negative_log_likelihood,
                    args=(X,y,numpy.array([1,regularization])))
    plt.suptitle('Regularization on w1 with $\lambda$ = '+ str(regular
ization))
    plot_data(X,y,res2.x,numpy.array([1,regularization]))
```

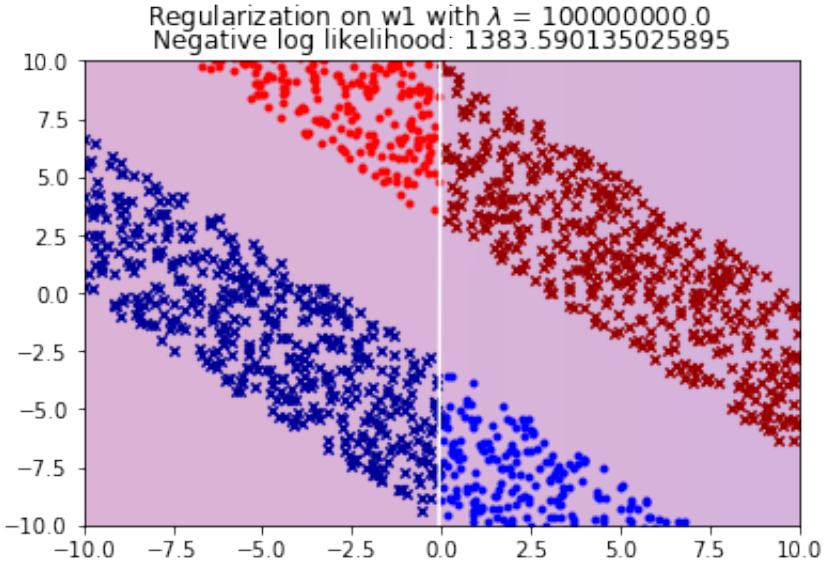

Regularization on w1 with λ = 1000.0
Negative log likelihood: 343.56315043868733
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0 −7.5 −5.0 −2.5 0.0 2.5 5.0 7.5 10.0


Regularization on w1 with λ = 10000.0
Negative log likelihood: 875.9681074071361
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0 −7.5 −5.0 −2.5 0.0 2.5 5.0 7.5 10.0


Regularization on w1 with λ = 100000.0
Negative log likelihood: 1288.0355824145513
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0 −7.5 −5.0 −2.5 0.0 2.5 5.0 7.5 10.0


Regularization on w1 with λ = 1000000.0
Negative log likelihood: 1373.1368490207142
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0 −7.5 −5.0 −2.5 0.0 2.5 5.0 7.5 10.0





observation:

- if we do not regularize on the logistic regression, the decision boundary would perfectly separate the two classes with no training error.
- if we regularize on w_1 , the training error increses as λ goes larger (penalize more on w_1). Eventually the decision boundary will be a horizontal line as parameter $w_1 = 0$.
- if we regularize on w_2 , as λ increases, the decision boundary becomes a vertical line as $w_2 = 0$.
- in both cases of regularization, training error increases as we penealize more on parameters.

In [ ]:

Problem 4: Decision Tree (35 points)

Decision trees are useful for classification. They follow a tree structure by splitting the data among the feature at each level which is most descriptive (gives largest information gain). For more information on decision trees see this chapter of Mitchell's machine learning book.

We will be classifying whether a banknote is fradulent or not using this Banknotes dataset. The dataset provides signal-based features of the banknotes and we will predict whether the banknote is fradulent (1) or not (0).

We will walk you through writing some key functions of a decision tree from scratch.

(a) (3 points) In the decision tree iPython notebook, start by importing the data (numpy array or pandas dataframe both work) and splitting the dataset by train and test (usually 80% train, 20% test). If you'd like to shuffle the data before splitting, you can use `numpy.random.shuffle`

(b) (8 points) Next, you will implement some functions in the `class DecisionTree`, which uses the `Node` class to build a decision tree. For each level, we calculate the current uncertainty, then split upon the feature at the threshold which gives us the highest information gain. First, implement the function `get_uncertainty` which implements the entropy and gini index uncertainty depending on the keyword `metric` that is given.

(c) (8 points) Using the `get_uncertainty` method, implement the `getInfoGain` method which calculates the information gain when splitting the data n `node.data` on `split_index`. `node.data` stores the relevant data at each node of the decision tree (root note will have all the data, then its children will split the data in 2 parts on the `split_index`, etc.)

(d) (10 points) Lastly, in `get_feature_threshold` find the feature that gives the largest information gain and **assign** the feature number (column number) to `node.feature` while updating the threshold values (check the docstrings for more detailed information).

(e) (6 points) After you implement the above steps, the `buildTree` function will recursively build the decision tree. Then when you run `homework_evaluate` which will call `self.predict` to evaluate the accuracy on the test set. Try different values of K (depth of tree a.k.a. number of features the tree will split on) and compare the performance. Which feature gives the largest information gain? Which feature is the least useful for the decision tree? How does varying the uncertainty metric (gini versus entropy) vary the performance (and why)? (For these questions we're looking for some analysis of the resulting model and some thought. Does not have to be a long paragraph.)

Congrats! You built a decision tree! Now (optional) try it on other data for fun. We've included a breast cancer dataset with both cardinal and categorical variables for you to try. With categorical variables, we one-hot encode them (e.g. for a 3-class categorical variable, 2 is represented as [0, 1, 0]) to allow the decision tree to classify them correctly (think about why this is the case, hint: does the magnitude of the number give us useful information here?).

Hello and welcome to decision trees!

Decision trees are often pretty effect learning algorithms, and certainly serve as an interesting technical exercise in data preprocessing and recursion. This notebook will walk you through some of the basic notions of how the implementation should be executed, and also give you a chance to write some of your own code.

Suggested reading before you start:
https://www.cs.princeton.edu/courses/archive/spring07/cos424/papers/mitchell-dectrees.pdf (https://www.cs.princeton.edu/courses/archive/spring07/cos424/papers/mitchell-dectrees.pdf)

Throughout the notebook you'll see TODO tags in the comments. This is where you should insert your own code to make the functions work! If you get stuck, we encourage you to come to office hours. You can also try to look at APIs and documentation online to try to get a sense how certain methods work. If you take inspiration from any source online other than official documentation, please be sure to cite the resource! Good luck!

In [64]:
```
import pandas as pd
import scipy.io
import numpy as np
```

We will be using decision trees to classify if a banknote is fradulent (class 1) or not fradulent (class 0). Download data from https://archive.ics.uci.edu/ml/datasets/banknote+authentication# (https://archive.ics.uci.edu/ml/datasets/banknote+authentication#)

Import data

In [65]:
```
# TODO YOUR CODE HERE
def import_data(split = 0.8, shuffle=False):
    """Read in the data, split it by split percentage into train and t
est data,
    and return X_train, y_train, X_test, y_test as numpy arrays"""
    # TODO
    df = pd.read_csv('./data_banknote_authentication.txt',
                     names = ['vairance','skewness','curtosis','entrop
y','class'])
    df_array = np.array(df)
    #shuffle data
    np.random.shuffle(df_array)
    train = df_array[:round(split*len(df_array))]
    test = df_array[round(split*len(df_array)):]
    X_train = train[:,0:4]
    y_train = train[:,-1].astype(int)
    X_test = test[:,0:4]
    y_test = test[:,-1].astype(int)
    print("data imported")

    return X_train, y_train, X_test, y_test
```

In [66]:
```
class Node:
    """Each node of our decision tree will hold values such as left an
d right children,
    the data and labels being split on, the threshold value & index in
the dataframe for a particular feature,
    and the uncertainty measure for this node"""
    def __init__(self, data, labels, depth):
        """
        data: X data
        labels: y data
        depth: depth of tree
        """
        self.left = None
        self.right = None

        self.data = data
        self.labels = labels
        self.depth = depth

        self.threshold = None # threshold value
        self.threshold_index = None # threshold index
        self.feature = None # feature as a NUMBER (column number)
        self.label = None # y label
        self.uncertainty = None # uncertainty value
```

In [67]:
```
class DecisionTree:
```

```
    def __init__(self, K=5, verbose=False):
        """
        K: number of features to split on
        """
        self.root = None
        self.K = K
        self.verbose = verbose

    def buildTree(self, data, labels, metric ='entropy'):
        """Builds tree for training on data. Recursively called _build
Tree"""
        self.root = Node(data, labels, 0)
        if self.verbose:
            print("Root node shape: ", data.shape, labels.shape)
        self._buildTree(self.root,  metric)

    def _buildTree(self, node,  metric ='entropy'):

        # get uncertainty measure and feature threshold
        node.uncertainty = self.get_uncertainty(node.labels)
        self.get_feature_threshold(node, metric)

        index = node.data[:, node.feature].argsort()  # sort feature f
or return
        node.data = node.data[index]
        node.labels = node.labels[index]

        # check label distribution.
        label_distribution = np.bincount(node.labels)
        majority_label = node.labels[0] if len(label_distribution) ==
1 else np.argmax(label_distribution)

        if self.verbose:
            print("Node uncertainty: %f" % node.uncertainty)

        # Split left and right if threshold is not the min or max of t
he feature or every point has the
        # same label.
        if node.threshold_index == 0 or node.threshold_index == node.d
ata.shape[0] or \
            len(label_distribution) == 1:
            node.label = majority_label
        else:
            node.left = Node(node.data[:node.threshold_index], node.la
bels[:node.threshold_index], node.depth + 1)
            node.right = Node(node.data[node.threshold_index:], node.l
abels[node.threshold_index:], node.depth + 1)
            node.data = None
            node.labels = None
```

```
            # If in last layer of tree, assign predictions
            if node.depth == self.K:
                if len(node.left.labels) == 0:
                    node.right.label = np.argmax(np.bincount(node.righ
t.labels))
                    node.left.label = 1 - node.right.label
                elif len(node.right.labels) == 0:
                    node.left.label = np.argmax(np.bincount(node.left.
labels))
                    node.right.label = 1 - node.left.label
                else:
                    node.left.label = np.argmax(np.bincount(node.left.
labels))
                    node.right.label = np.argmax(np.bincount(node.righ
t.labels))
                return

            else: # Otherwise continue training the tree by calling _b
uildTree
                self._buildTree(node.left)
                self._buildTree(node.right)

    def predict(self, data_pt):
        return self._predict(data_pt, self.root)

    def _predict(self, data_pt, node):
        feature = node.feature
        threshold = node.threshold
        if node.label is not None:
            return node.label
        elif data_pt[node.feature] < node.threshold:
            return self._predict(data_pt, node.left)
        elif data_pt[node.feature] >= node.threshold:
            return self._predict(data_pt, node.right)

    def get_feature_threshold(self, node, metric ='entropy'):
        """ TODO
        Find the feature that gives the largest information gain.
        Update node.threshold, node.threshold_index, and node.feature
(a number representing the feature. e.g. 2nd column feature would be 1
)
        Make sure to sort the columns of data before you try to find t
he threshold index (look at numpy argsort) and set the values
        for node.threshold, node.threshold_index, and node.feature
        return: None
        """
        node.threshold = 0
        node.threshold_index = 0
        node.feature = 0
        # TODO YOUR CODE HERE
```

```
        n = node.data.shape[0]
        n_feature = node.data.shape[1]
        info_gain = 0
        for i in range(n_feature):
            sort = np.argsort(node.data[:,i])
            node.data = node.data[sort]
            node.labels = node.labels[sort]
            for j in range(n-1):
                gain = self.getInfoGain(node, j+1, metric)
                if info_gain == 0:
                    info_gain = gain
                    node.threshold = node.data[j+1,i]
                    node.threshold_index = j+1
                    node.feature = i
                elif gain > info_gain:
                    info_gain = gain
                    node.threshold = node.data[j+1,i]
                    node.threshold_index = j+1
                    node.feature = i




    def getInfoGain(self, node, split_index, metric ='entropy'):
        """
        TODO Get information gain using the variables in the parameter
s, \
        split_index: index in the feature column that you are splittin
g the classes on
        return: information gain (float)
        """
        # TODO YOUR CODE HERE
        Q = self.get_uncertainty(node.labels, metric)
        Q_r = self.get_uncertainty(node.labels[0:split_index], metric)
        Q_l = self.get_uncertainty(node.labels[split_index:], metric)
        gain = Q - (Q_l * (split_index)/node.labels.shape[0]
                    + Q_r * (1-(split_index/node.labels.shape[0])))
        return gain

    def get_uncertainty(self, labels, metric="entropy"):
        """
        TODO Get uncertainty. Implement entropy AND gini index metrics
.
        np.bincount(labels) and labels.shape might be useful here
        return: uncertainty (float)
        """

        if labels.shape[0] == 0:
            return 1
```

```
        # TODO YOUR CODE HERE

        else:
            Q = 0
            p_mk = (1/labels.shape[0]) * np.bincount(labels)[0] #np.bi
ncount(labels)[0] output majority of type
            if metric == 'entropy':
                if p_mk !=1 and p_mk !=0:
                    Q = p_mk * np.log(p_mk) + (1-p_mk) * np.log(p_mk)
            if metric == 'gini':
                Q = p_mk*(1-p_mk)*2
        return Q




    def printTree(self):
        """Prints the tree including threshold value and feature name"
""
        self._printTree(self.root)

    def _printTree(self, node):
        if node is not None:
            if node.label is None:
                print("\t" * node.depth, "(%d, %d)" % (node.threshold,
node.feature))
            else:
                print("\t" * node.depth, node.label)
            self._printTree(node.left)
            self._printTree(node.right)

    def homework_evaluate(self, X_train, labels, X_test, y_test):
        n = X_train.shape[0]

        count = 0
        for i in range(n):
            if self.predict(X_train[i]) == labels[i]:
                count += 1

        print("The decision tree is %d percent accurate on %d training
data" % ((count / n) * 100, n))

        n = X_test.shape[0]

        count = 0
        for i in range(n):
            if self.predict(X_test[i]) == y_test[i]:
                count += 1

        print("The decision tree is %d percent accurate on %d test dat
```

```
a" % ((count / n) * 100, n))

        return count / n
```

Run the tree

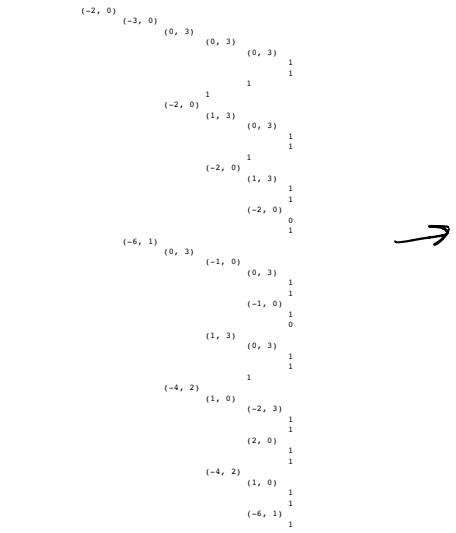
Try different values of K (depth of tree a.k.a. number of features the tree will split on) and compare the performance. Which feature gives the largest information gain? Which feature is the least useful for the decision tree?

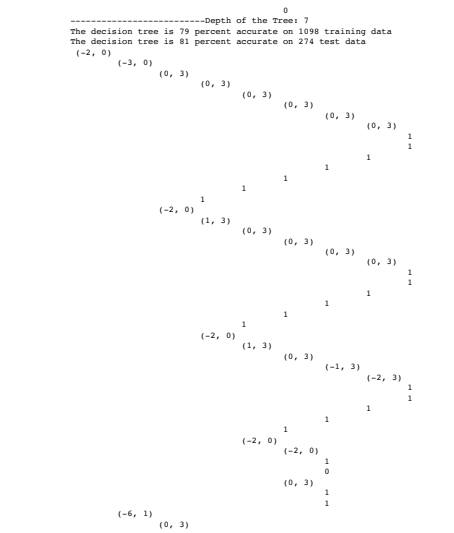
In [68]:
```
X_train, y_train, X_test, y_test = import_data(split=0.8)
```

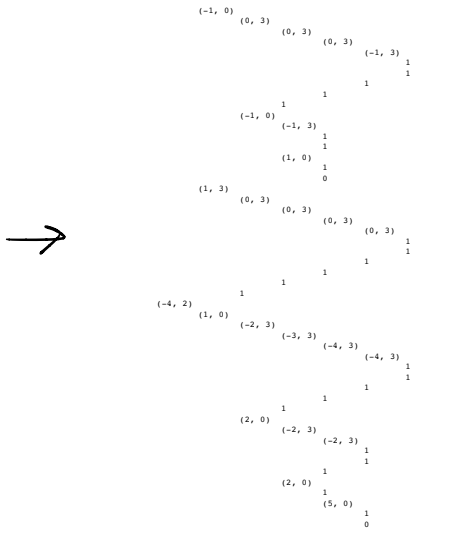
```
data imported
```

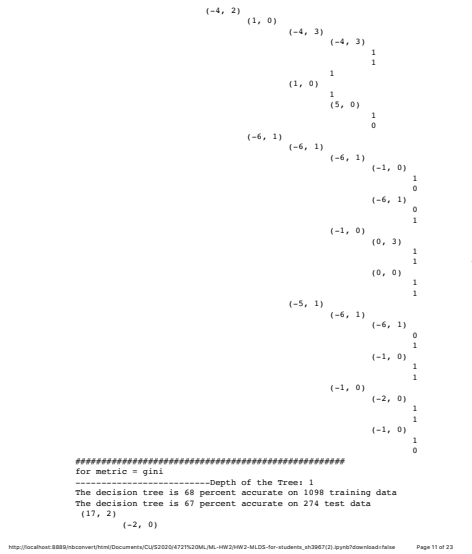
In [70]:
```
print('for metric = entropy')
for k in range(1,10,3):
    print('-------------------------Depth of the Tree:', k)
    tree = DecisionTree(K=k, verbose=False)
    tree.buildTree(X_train, y_train, metric = 'entropy')
    tree.homework_evaluate(X_train, y_train, X_test, y_test)
    tree.printTree()
print('#####################################################')

print('for metric = gini')
for k in range(1,10,3):
    print('-------------------------Depth of the Tree:', k)
    tree = DecisionTree(K=k, verbose=False)
    tree.buildTree(X_train, y_train, metric = 'gini')
    tree.homework_evaluate(X_train, y_train, X_test, y_test)
    tree.printTree()
```

```
for metric = entropy
-------------------------Depth of the Tree: 1
The decision tree is 72 percent accurate on 1098 training data
The decision tree is 71 percent accurate on 274 test data
 (-2, 0)
         (-3, 0)
                 1
                 1
         (-6, 1)
                 1
                 0
-------------------------Depth of the Tree: 4
The decision tree is 75 percent accurate on 1098 training data
The decision tree is 74 percent accurate on 274 test data
```

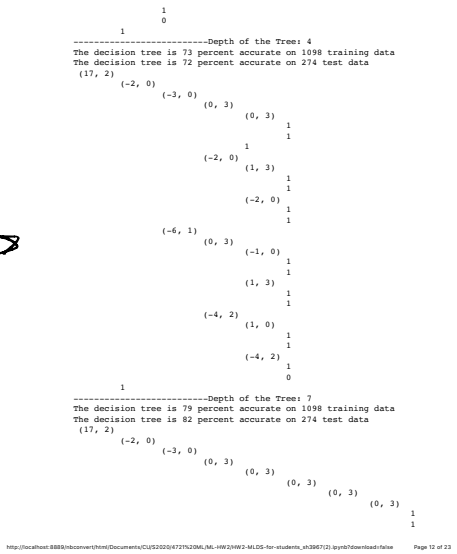




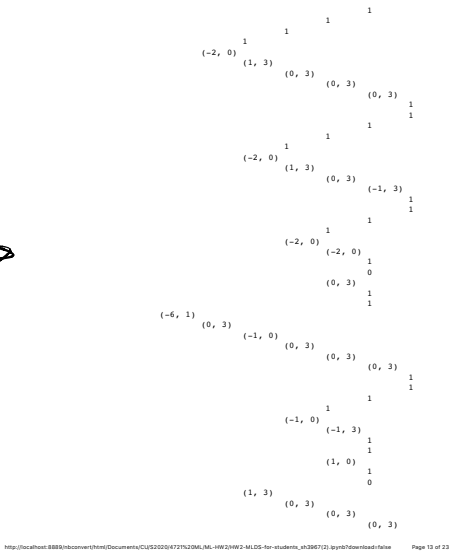


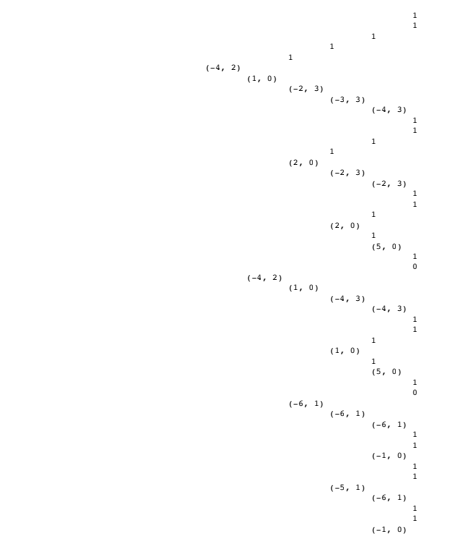


```
##################################################
for metric = gini
--------------------------Depth of the Tree: 1
The decision tree is 68 percent accurate on 1098 training data
The decision tree is 67 percent accurate on 274 test data
 (17, 2)
        (-2, 0)
```

```
--------------------------Depth of the Tree: 7
The decision tree is 79 percent accurate on 1098 training data
The decision tree is 82 percent accurate on 274 test data
 (17, 2)
        (-2, 0)
                (-3, 0)
                        (0, 3)
                                (0, 3)
```



I put them together b/c the trees are large

1
0

1

- For both metric, the accuracy of decision tree model increase by adding more depth.
- By observing printed tree, 0 is the feature appears most frequently, while 3 is the least frequent feature. Therefore, feature 0 provides largest information gain. Feature 3 is the least useful for the tree.
- Model performance does't vary a lot provided different uncertainty metric, `gini` and `entropy`. The reason is that both of the two metric are measuring impurity with similar methods. The two function change in same direction. Therefore, the output is similar

---

Optional Decision Tree Exercise

This section is designed to give you some exposure to typical preprocessing and allow you to run your decision tree code on another example.

All of the preprocessing has been done for you — there's nothing you need to fill in, but it may be worthwhile to tinker with some of the pieces to make sure you understand how everything fits together.

Otherwise, if your decision tree code works on the bank notes example, you should be able to run through this straight away.

Step 1: Data Preprocessing

To start, you'll need to download the data files from [https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer/ (https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer/)](https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer/). The file `breast-cancer.data` contains the actual data you'll need and the file `breast-cancer.names` gives some information about the researchers and the data types (it's probably worth looking at to give you a sense of what's going on).

Note: If you try to open `breast-cancer.data` or `breast-cancer.names` directly, your computer might not know how to handle the file format. To view them, you need to change the file types from .data and .names to .txt — this can be accomplished by simply changing the file name from `breast-cancer.data` to `breast-cancer.txt`

Now let's load the data file into the notebook. All you need to do is put it in the same directory as the notebook and the cell below should locate the appropriate file.

In [48]:
```
import os
current_directory = os.listdir()

try:
  cancer_files = [file for file in current_directory if 'cancer' in fi
le]
  data_file = [file for file in cancer_files if 'names' not in file][0
]
except IndexError:
  print('The breast cancer files were not found. Please upload again.'
)
  data_file = None

if data_file:
  print(f'The data file has been located as {data_file}.')
```

```
The data file has been located as breast-cancer.data.
```

!! If the cell above returned a file that you don't recognize as the correct data file, you must go back to the upload step and ensure you have successfully uploaded your files before proceeding. !!

Now we'll try to read the sample into our environment:

In [49]:
```
samples = []

with open(data_file, 'r') as file:
  samples = file.readlines()

samples = [sample.split(',') for sample in samples]

print(f'There are {len(samples)} samples in the dataset')
```

```
There are 286 samples in the dataset
```

We can now start defining how we want to map our sample values to numeric features that can be used in the decision tree algorithm below. Let's first store our samples in a pandas DataFrame so that it'll be easier to view and work with.

In [50]:
```
import pandas as pd
import numpy as np

columns = ['class', 'age', 'menopause', 'tumor-size', 'inv-nodes',
           'node-caps', 'deg-malig', 'breast', 'breast-quad', 'irradia
t']

samples_df = pd.DataFrame(samples, columns=columns)
samples_df['irradiat'] = samples_df['irradiat'].str.replace('\\n', '')

samples_df
```

Out[50]:

	class	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat
0	no-recurrence-events	30-39	premeno	30-34	0-2	no	3	left	left_low	no
1	no-recurrence-events	40-49	premeno	20-24	0-2	no	2	right	right_up	no
2	no-recurrence-events	40-49	premeno	20-24	0-2	no	2	left	left_low	no
3	no-recurrence-events	60-69	ge40	15-19	0-2	no	2	right	left_up	no
4	no-recurrence-events	40-49	premeno	0-4	0-2	no	2	right	right_low	no
...	...	...	...	...	...	...	...	...	...	...
281	recurrence-events	30-39	premeno	30-34	0-2	no	2	left	left_up	no
282	recurrence-events	30-39	premeno	20-24	0-2	no	3	left	left_up	yes
283	recurrence-events	60-69	ge40	20-24	0-2	no	1	right	left_up	no
284	recurrence-events	40-49	ge40	30-34	3-5	no	3	left	left_low	no
285	recurrence-events	50-59	ge40	30-34	3-5	no	3	left	left_low	no

286 rows × 10 columns

Our DataFrame looks great! Now that we have the raw data stored in an interpretable way, it's good practice to make a copy before trying to encode everything — if something goes wrong, it'll be easy to just come back up here and reset the encoded DataFrame.

In [51]:
```
encoded_samples_df = samples_df.copy()
```

In the following few cells, we're going to define the different types of variables that exist in our dataset and make sure we assign the appropriate type of encoding.

To start, we'll define our binary variables (0,1) and create Python dictionaries to map the string labels to binary integer values.

It's also helpful to store all of the columns and dictionaries in two lists so that we can easily reference them later on.

In [52]:
```
binary_cols = ['class', 'breast', 'irradiat']

class_map = {'no-recurrence-events': 0, 'recurrence-events': 1}
breast_map = {'left': 0, 'right': 1}
irrad_map = {'yes': 1, 'no': 0}

binary_maps = [class_map, breast_map, irrad_map]
```

Next, we've defined our ordinal variables (those that have obvious ordered structure). Instead of hard-coding the maps here, it's nice to have a function that can take any number of ordinal columns and instantly create all of the corresponding maps. The code below does that exactly, relying on the integer value of the first number in the range (i.e. '50-65') as the sorting key.

In [53]:
```
ordinal_cols = ['age', 'tumor-size', 'inv-nodes', 'deg-malig']

def first_val(x):
  return eval(x.split('-')[0].replace("'", ''))

ordinal_maps = {}

for col in ordinal_cols:
  uniques = sorted(list(encoded_samples_df[col].unique()), key=first_v
al)
  val_map = {}
  i = 1

  for val in uniques:
    val_map[val] = i
    i += 1

  ordinal_maps[col] = val_map
```

Finally, we have several columns that take n > 2 discrete values. Binary encoding won't work here, so we'll use one-hot encoding. Just like in the cell immediately above, we'll use a function here to take an arbitrary number of columns and encode their unique values as one-hot vectors.

In [54]:
```
one_hot_cols = ['menopause', 'breast-quad', 'node-caps']
one_hot_maps = {}

def one_hot_map(col, df):
  one_hot = {}
  unique_vals = list(df[col].unique())
  one_hot_vecs = np.identity(len(unique_vals))

  for i in range(len(unique_vals)):
    one_hot[unique_vals[i]] = one_hot_vecs[i]

  return one_hot

for col in one_hot_cols:
  one_hot_maps[col] = one_hot_map(col, encoded_samples_df)
```

Now that we have all of our encoding maps set up, we'll zip them all up into a master dictionary so that we can easily iterate through them on our `encoded_samples_df` DataFrame.

In [55]:
```
all_maps = dict(one_hot_maps, **ordinal_maps)
for col, val_map in zip(binary_cols, binary_maps):
  all_maps[col] = val_map
```

Before proceeding, let's double check to make sure you've captured all of the columns:

In [56]:
```
if len(all_maps) != len(columns):
  print("You're missing a map! Go back and double check that you didn'
t miss a column.")
else:
  print("Looks like you successfully created all the maps! Nice work!"
)
```

```
Looks like you successfully created all the maps! Nice work!
```

And now the moment of truth! Let's apply all of our encoding maps on the DataFrame to turn everything into useable features for our decision trees algorithm.

In [57]:
```
for col in columns:
  encoded_samples_df[col] = encoded_samples_df[col].map(all_maps[col])

encoded_samples_df
```

Out[57]:

	class	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat
0	0	2	[1.0, 0.0, 0.0]	7	1	[1.0, 0.0, 0.0]	3	0	[1.0, 0.0, 0.0, 0.0, 0.0, 0.0]	0
1	0	3	[1.0, 0.0, 0.0]	5	1	[1.0, 0.0, 0.0]	2	1	[0.0, 1.0, 0.0, 0.0, 0.0, 0.0]	0
2	0	3	[1.0, 0.0, 0.0]	5	1	[1.0, 0.0, 0.0]	2	0	[1.0, 0.0, 0.0, 0.0, 0.0, 0.0]	0
3	0	5	[0.0, 1.0, 0.0]	4	1	[1.0, 0.0, 0.0]	2	1	[0.0, 0.0, 1.0, 0.0, 0.0, 0.0]	0
4	0	3	[1.0, 0.0, 0.0]	1	1	[1.0, 0.0, 0.0]	2	1	[0.0, 0.0, 0.0, 1.0, 0.0, 0.0]	0
...	...	...	...	...	...	...	...	...	...	...
281	1	2	[1.0, 0.0, 0.0]	7	1	[1.0, 0.0, 0.0]	2	0	[0.0, 0.0, 1.0, 0.0, 0.0, 0.0]	0
282	1	2	[1.0, 0.0, 0.0]	5	1	[1.0, 0.0, 0.0]	3	0	[0.0, 0.0, 1.0, 0.0, 0.0, 0.0]	1
283	1	5	[0.0, 1.0, 0.0]	5	1	[1.0, 0.0, 0.0]	1	1	[0.0, 0.0, 1.0, 0.0, 0.0, 0.0]	0
284	1	3	[0.0, 1.0, 0.0]	7	2	[1.0, 0.0, 0.0]	3	0	[1.0, 0.0, 0.0, 0.0, 0.0, 0.0]	0
285	1	4	[0.0, 1.0, 0.0]	7	2	[1.0, 0.0, 0.0]	3	0	[1.0, 0.0, 0.0, 0.0, 0.0, 0.0]	0

286 rows × 10 columns

Double check your `encoded_samples_df` here to make sure everything passes the sniff test!

Assuming it all looks good, the final step is to break up our DataFrame into individual samples, unpack the nested arrays into their constituent values, and concatenate everything together into a final sample vector...

In [58]:
```
sample_list = list(encoded_samples_df.to_numpy())
```

In [59]:
```
def unpack_nested_arrays(vec):
  final_vec = np.array([])

  for e in vec:
    if isinstance(e, int):
      e = np.array([e])
    final_vec = np.concatenate((final_vec, e))

  return final_vec
```

In [60]:
```
final_vecs = []
for e in sample_list:
  final_vecs.append(unpack_nested_arrays(e))

final_vecs = np.array(final_vecs, dtype=int)
```

Great! The `final_vecs` variable should now point to a 286x19 numpy.ndarray containing all of our samples. Let's finally move on to the machine learning!

Step 2: Decision Tree

We'll run the decision tree process here again.

In [61]:
```
def import_cancer_data(split=0.8, shuffle=True, CUTOFF=0, bins = 256):

    if shuffle:
      np.random.shuffle(final_vecs)

    num_samples = final_vecs.shape[0]
    num_train_samples = int(num_samples*split)

    train_data, test_data = final_vecs[:num_train_samples, :], final_v
ecs[num_train_samples:, :]

    X_train = train_data[:, 1:]
    y_train = train_data[:, 0]
    X_test = test_data[:, 1:]
    y_test = test_data[:, 0]

    print(X_train.shape)
    print(y_train.shape)

    return X_train, y_train, X_test, y_test
```

Here's the final fit/evaluation code again, but this time using the breast cancer data. You should get somewhere around 0.75 accuracy for the decision tree on this dataset — not 100%, but not too bad for the humble decision tree!

In [62]:
```
X_train, y_train, X_test, y_test = import_cancer_data(split=0.8)

tree = DecisionTree(K=3, verbose=False)
tree.buildTree(X_train, y_train)

tree.homework_evaluate(X_train, y_train, X_test, y_test)
```

```
(228, 18)
(228,)
The decision tree is 63 percent accurate on 228 training data
The decision tree is 77 percent accurate on 58 test data
```

Out[62]:
```
0.7758620689655172
```

In [ ]:

In [ ]:

In [ ]: