Homework III-KEY

1. (Point Estimation) You are given a coin and a thumbtack and you put Beta priors Beta(100; 100) and Beta(1; 1) on the coin and thumbtack respectively. You perform the following experiment: toss both the thumbtack and the coin 100 times. To your surprise, you get 60 heads and 40 tails for both the coin and the thumbtack. Are the following two statements true or false?
   _ The MLE estimate of both the coin and the thumbtack is the same but the MAP estimate is not.
   _ The MAP estimate of the parameter θ (probability of landing heads) for the coin is greater than the MAP estimate of θ for the thumbtack.
   Explain your answer mathematically. [5 Points]

The first statement is True. MLE estimate = 6/10 for both cases. The MAP estimates are not because the beta priors are different.
The second statement is False. MAP estimate for the coin

$$\frac{\boldsymbol{\alpha_H + \beta_H - 1}}{\boldsymbol{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}}$$

=(60 + 100 -1) / (60 + 100 + 40 + 100 - 2)
=159/298
= 0.53356
MAP estimate for the thumbtack
= (60 + 1-1) / (60 + 1 + 40 + 1-2)
=60/100
= 0.6

2. Consider learning a function X -> Y where Y is boolean, where X $\langle X_1, X_2 \rangle$, and where X_1 is a boolean variable and X_2 is continuous variable. State the parameters that must be estimated to define a Naïve Bayes classifier in this case. Give the formula for computing P(Y|X), in terms of these parameters and the feature values X_1 and X_2. [5 Points]

7 parameters two for Boolean X_1 (one for $P(X_1|Y=1)$ and one for $P(X_1|Y=0)$) and 4 for Gaussian X_2 (2 for $P(X_1|Y=1)$ and 2 for $P(X_1|Y=0)$). And one more for $P(Y)$.

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)}$$

$$= \frac{P(X_1,\ \ X_2|Y) * P(Y)}{P(X)}$$

$$= \frac{P(X_1|Y) * \ \ P(X_2|Y) * P(Y)}{P(X)}$$

3. Naïve Bayes

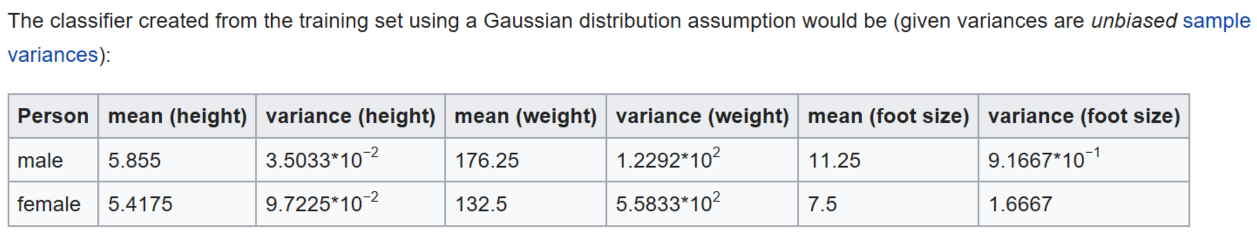
   Classify whether a given person is a male or a female based on the measured features using naïve bayes classifier. The features include height, weight, and foot size.

Training Data for the classifier is given in the below table.

Person	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

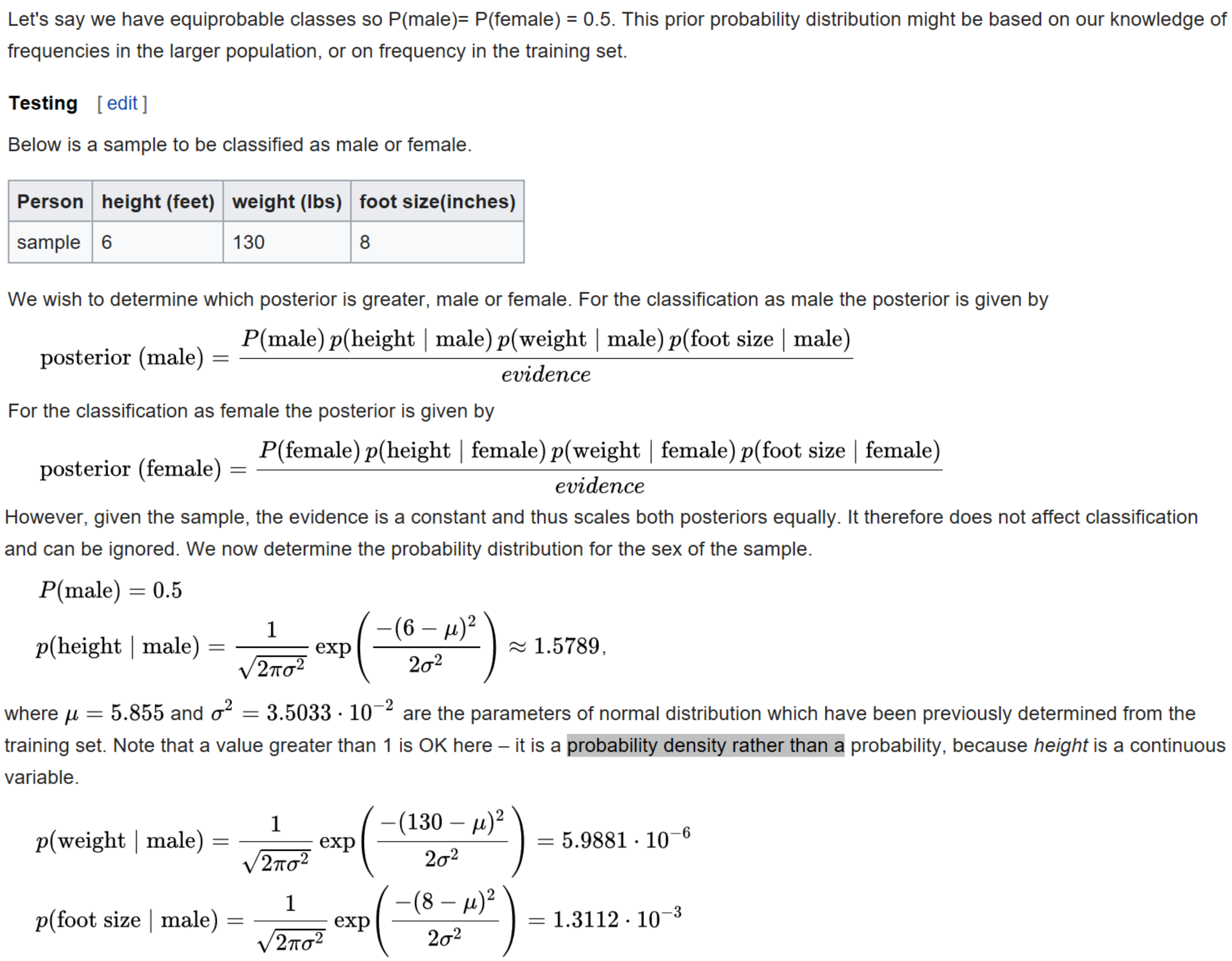
Below is a sample to be classified as male or female.

Person	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8



The classifier created from the training set using a Gaussian distribution assumption would be (given variances are *unbiased* sample variances):

Person	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	$3.5033*10^{-2}$	176.25	$1.2292*10^{2}$	11.25	$9.1667*10^{-1}$
female	5.4175	$9.7225*10^{-2}$	132.5	$5.5833*10^{2}$	7.5	1.6667



Let's say we have equiprobable classes so P(male)= P(female) = 0.5. This prior probability distribution might be based on our knowledge of frequencies in the larger population, or on frequency in the training set.

Testing [edit]

Below is a sample to be classified as male or female.

Person	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

We wish to determine which posterior is greater, male or female. For the classification as male the posterior is given by

$$\text{posterior (male)} = \frac{P(\text{male})\, p(\text{height} \mid \text{male})\, p(\text{weight} \mid \text{male})\, p(\text{foot size} \mid \text{male})}{evidence}$$

For the classification as female the posterior is given by

$$\text{posterior (female)} = \frac{P(\text{female})\, p(\text{height} \mid \text{female})\, p(\text{weight} \mid \text{female})\, p(\text{foot size} \mid \text{female})}{evidence}$$

However, given the sample, the evidence is a constant and thus scales both posteriors equally. It therefore does not affect classification and can be ignored. We now determine the probability distribution for the sex of the sample.

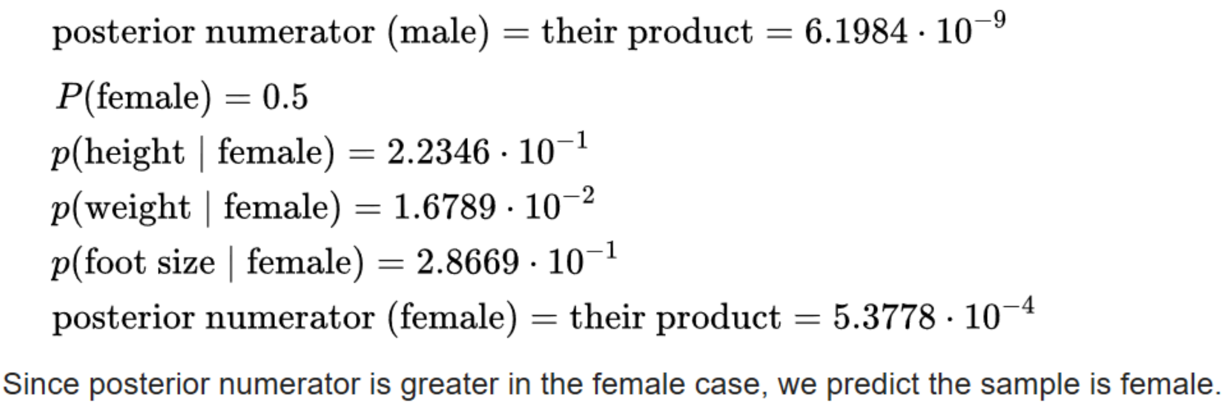
$$P(\text{male}) = 0.5$$

$$p(\text{height} \mid \text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6-\mu)^2}{2\sigma^2}\right) \approx 1.5789,$$

where $\mu = 5.855$ and $\sigma^2 = 3.5033 \cdot 10^{-2}$ are the parameters of normal distribution which have been previously determined from the training set. Note that a value greater than 1 is OK here – it is a probability density rather than a probability, because *height* is a continuous variable.

$$p(\text{weight} \mid \text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(130-\mu)^2}{2\sigma^2}\right) = 5.9881 \cdot 10^{-6}$$

$$p(\text{foot size} \mid \text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(8-\mu)^2}{2\sigma^2}\right) = 1.3112 \cdot 10^{-3}$$



$$\text{posterior numerator (male)} = \text{their product} = 6.1984 \cdot 10^{-9}$$

$$P(\text{female}) = 0.5$$

$$p(\text{height} \mid \text{female}) = 2.2346 \cdot 10^{-1}$$

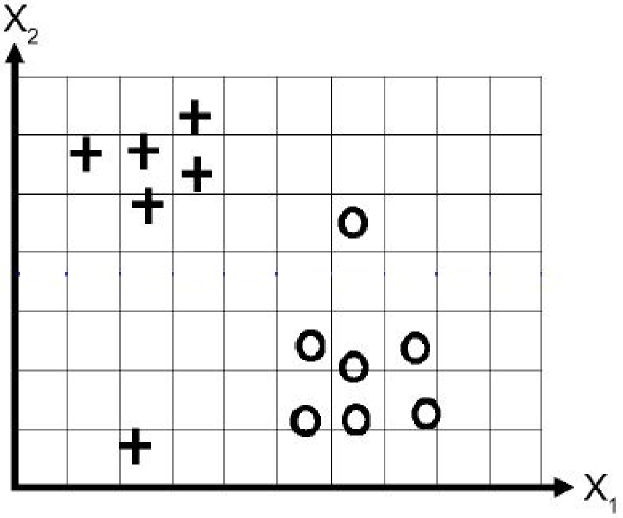
$$p(\text{weight} \mid \text{female}) = 1.6789 \cdot 10^{-2}$$

$$p(\text{foot size} \mid \text{female}) = 2.8669 \cdot 10^{-1}$$

$$\text{posterior numerator (female)} = \text{their product} = 5.3778 \cdot 10^{-4}$$

Since posterior numerator is greater in the female case, we predict the sample is female.

4. Regularization separate terms in 2d logistic regression



a. Consider the data in Figure where we fir model $p(y = 1 \mid x, w) = \sigma(\mathrm{w}_0 + \mathrm{w}_1\mathrm{x}_1 + \mathrm{w}_2\mathrm{x}_2)$. Suppose we fir the model by maximum likelihood or we minimize $J(w) = -l(w, D_{train})$

where $l(w, D_{train})$ is the log likelihood on the training set. Sketch a possible decision boundary corresponding to w. Is your answer (decision boundary) unique? How many classification errors does your method make on the training set?

b. Now suppose we regularize only the w_0 parameter i.e. we minimize

$$J(w) = -l(w, D_{train}) + \lambda w_0^2$$

Suppose λ is a very large number, so we regularize w_0 all the way to 0, but all other parameters are unregularized. Sketch a possible decision boundary. How many classification errors does your method make on training set?

c. Now suppose we regularize only the w_1 parameter i.e. we minimize

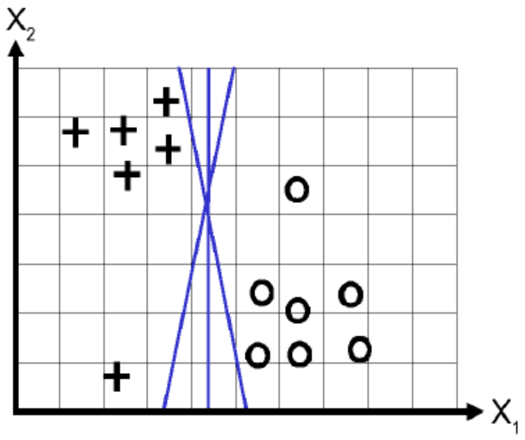
$$J(w) = -l(w, D_{train}) + \lambda w_1^2$$

Sketch a possible decision boundary. How many classification errors does your method make on training set?

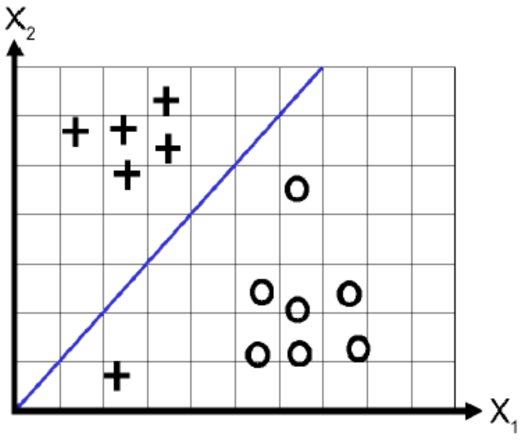
d. Now suppose we regularize only the w_2 parameter i.e. we minimize

$$J(w) = -l(w, D_{train}) + \lambda w_2^2$$

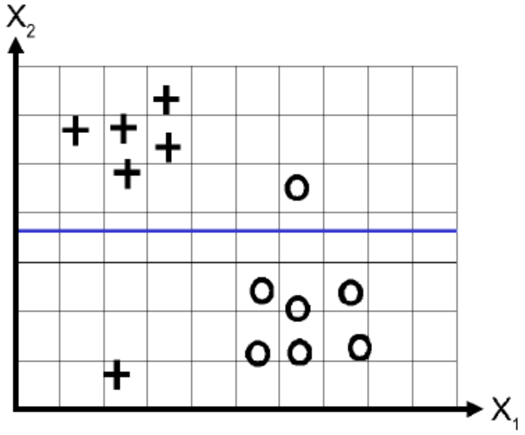
Sketch a possible decision boundary. How many classification errors does your method make on training set?



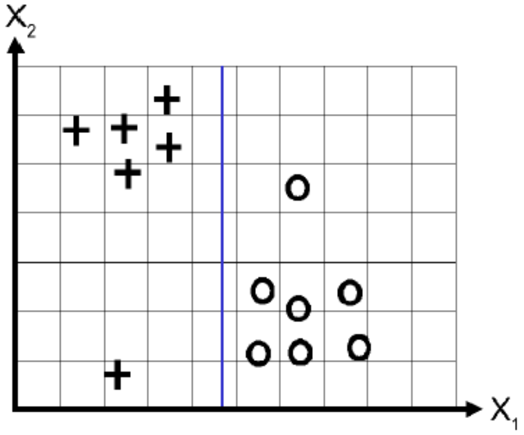
(a)



(b)



(c)



(d)

5. Using the following 2-D dataset (x1 and x2 are the attributes and y is the class variable), find the linear SVM classifier. Do your optimization using the dual problem. Namely, provide an explicit expression for the dual optimization problem, solve it (compute the values of the various α_i's) and use the solution to compute the weights attached to the two attributes as well as the bias term.

[10 Points]

Dataset:

X_1	X_2	Y
1	0	+
-1	2	-
0	-1	+

Since there are two points there will be two Lagragian multipliers one associated with each point.

Recall the dual formula

$$L = \sum_i \alpha_i \; - \frac{1}{2}\sum_i \sum_j \alpha_i \, \alpha_j \; y_i \, y_j \, x_i \, x_j$$

Calculate the dot product of all points. You will be doing vector multiplication

$$1 \quad 0 * \begin{matrix} 1 \\ 1 \end{matrix} \; = \; 1 * 1 + 0 * 1 = 1$$

	P1	P2	P3
P1	1	-1	0
P2	-1	5	-2
P3	0	-2	1

Substituting the dot products and y- values we get..

$$L = \; \alpha_1 + \; \alpha_2 + \; \alpha_3 - \frac{1}{2} \, \alpha_1^2 \, (1)(\,1) - \frac{1}{2}\alpha_1\alpha_2(-1)(-1) \; - \frac{1}{2}\alpha_1\alpha_3(1)(0) - \frac{1}{2}\alpha_1\alpha_2(-1)(-1) - \frac{1}{2} \, \alpha_2^2 \, (1)(\,5) \; - \frac{1}{2}\alpha_2\alpha_3(-1)(-2) \; - \frac{1}{2}\alpha_3\alpha_1(1)(0) \; - \frac{1}{2}\alpha_3\alpha_2(-1)(-2) \; - \frac{1}{2} \, \alpha_3^2 \, (1)(\,1)$$

$$L = \; \alpha_1 + \; \alpha_2 \; + \; \alpha_3 - \; \frac{1}{2}\alpha_1^2 - \alpha_1\alpha_2 - \frac{5}{2} \; \alpha_2^2 - 2\alpha_2\alpha_3 \; - \frac{1}{2} \; \alpha_3^2$$

We already know what we get when we differentiate L w.r.t to w and b

$$w = \sum_i \alpha_i y_i x_i$$

$$\sum_i \alpha_i y_i \; = 0$$

By using the second equation we get $\alpha_1 - \; \alpha_2 + \alpha_3 \; = 0,$ call it eq(1).

By differentiating L w.r.t to α_1, α_2 and α_3, we get following three equations

$$1 - \alpha_1 - \; \alpha_2 \; = 0$$

$$1 - \alpha_1 - \; 5\alpha_2 - \; 2\alpha_3 \; = 0$$

$$1 - 2\alpha_2 - \; \alpha_3 \; = 0$$

Call it eq(2), eq(3) and eq(4).

eq(2) and eq(3) contradict each other so we will use only eq(1), eq(2) and eq(4) to solve for all αs. so we

$$\alpha_1 = \frac{1}{2} \quad \alpha_2 = \frac{1}{2} \quad \alpha_3 = \; 0$$

By substituting the α_1 and α_2 in w equation we get

$$\boldsymbol{w} = \begin{array}{c} \frac{1}{2}(+1)(1) + \frac{1}{2}(-1)(-1) \\ \frac{1}{2}(+1)(0) + \frac{1}{2}(-1)(2) \end{array} = \begin{array}{c} 1 \\ -1 \end{array}$$

Substitute w in any one of the initial conditions i.e.
$\boldsymbol{w.x + b = +1\ for\ positive\ support\ vector\ and\ w.x + b = -1\ for\ negative}$

we will get b = 0 and b = 2 so average is b=1.

1. A SVM is trained with the following data: **[10 Points]**

x_1	x_2	class
-1	-1	-1
1	1	1
0	2	1

Let $\alpha_1, \alpha_2\ and\ \alpha_3$ be the lagrangian multipliers for the three data points.

a. Using polynomial kernel of degree 2 what (dual) optimization problem needs to be solved in terms of the lagrangian multipliers in order to determine their values? The polynomial kernel of degree d is given by the equation

$$K(x_i, x_j) = (1 + x_i^T x_j)^d$$

$$where\ x_i\ and\ x_j\ are\ input\ vectors$$

$$\boldsymbol{maximize \sum_{i=1}^{3} \alpha_i - \frac{1}{2} \sum_{i=1}^{3} \sum_{j=1}^{3} \alpha_i \alpha_j y_i y_j (1 + x_i^T x_j)^2}$$

subject to the constraint : $\boldsymbol{\alpha_1, \alpha_2, \alpha_3 \geq 0}$ and $\boldsymbol{-\alpha_1 + \alpha_2 + \alpha_3 = 0}$. The quantity $\boldsymbol{y_i y_j (1 + x_i^T x_j)^2}$ for different values of $\boldsymbol{i}$ and $\boldsymbol{j}$ is given by the cells of the following matrix

$$\begin{pmatrix} \mathbf{9} & \mathbf{-1} & \mathbf{-1} \\ \mathbf{-1} & \mathbf{9} & \mathbf{9} \\ \mathbf{-1} & \mathbf{9} & \mathbf{25} \end{pmatrix}$$

b. Let us say that we have solved the optimization problem and found that $\alpha_1 = \alpha_2 = \frac{1}{8} and\ \alpha_3 = 0$. Moreover b = 0. Can you tell me which of the data points are support vectors? Explain your answer.

The support vector are points 1 and 2 because $\boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_2}$ are greater than zero.

c. Assuming $\alpha_1 = \alpha_2 = \frac{1}{8}$, $\alpha_3 = 0$ and b = 0, how will the SVM classify the point (x_1 = -1, x_2 = 0)? Explain your answer?

The dot product of the kernel with the test point is (4, 0).

$$-\frac{\mathbf{1}}{\mathbf{8}}(\mathbf{4}) + \frac{\mathbf{1}}{\mathbf{8}}(\mathbf{0}) < \mathbf{0}$$

Therefore, the class is −1.

d. Assuming $\alpha_1 = \alpha_2 = \frac{1}{8}$, $\alpha_3 = 0$ and b = 0, how will the SVM classify the point (x_1 = 1, x_2 = 0)? Explain your answer?

The dot product of the kernel with the test point is (0, 4).

$$-\frac{\mathbf{1}}{\mathbf{8}}(\mathbf{0}) + \frac{\mathbf{1}}{\mathbf{8}}(\mathbf{4}) > \mathbf{0}$$

Therefore, the class is +1.