ECE M146 Introduction to Machine Learning

Lecture 11 - Spring 2021

Prof. Lara Dolecek

ECE Department, UCLA

Today's Lecture

Recap:

- Classification methods so far

New topics:

- Generative modeling
- Naïve Bayes Classifier

Recap – modeling classification problems

- We have previously studied parametric methods with a discriminant function.
- What were these methods ?

perceptron

SVM

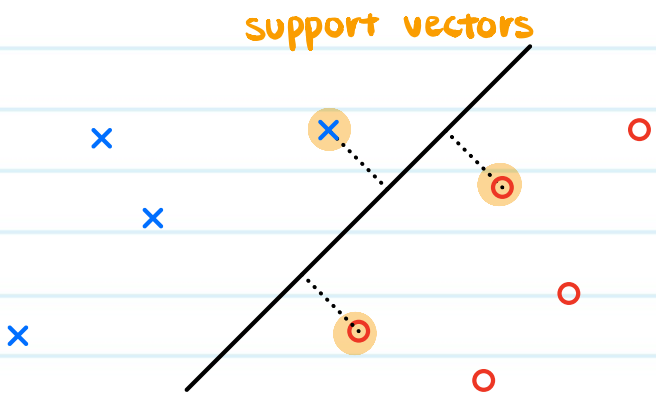
- We have also studied a method with discriminative modeling.
- What was this method ?

logistic regression

Last two lectures: SVM

- Key idea

maximize margin



qualitatively similar loss functions ==> soft SVM
and logistic regression perform similarly

Connections with Logistic Regression

$y_n = w^T x_n + b$

soft SVM

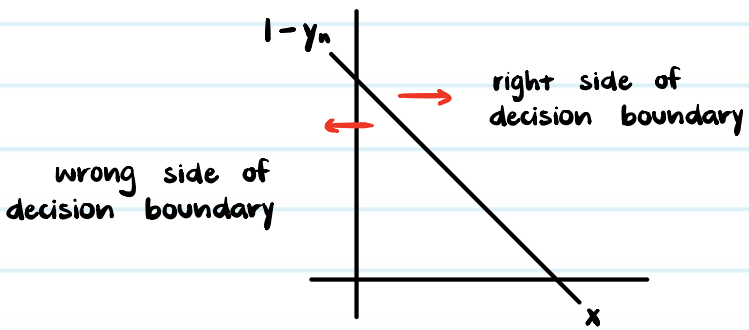
$t_n y_n = t_n (w^T x_n + b)$

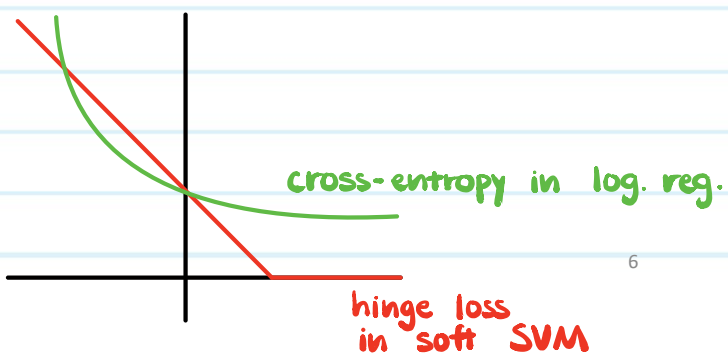
$t_n \in \{-1, +1\}$

$\varepsilon_n = 1 - t_n y_n$

$\min \frac{1}{2}\|w\|^2 + C \cdot \sum_{n=1}^{N} [1 - t_n y_n]_+$

notation: $[a]_+ = \begin{cases} a & \text{if } a \geq 0 \\ 0 & \text{else} \end{cases}$





Today's Lecture

Recap:

- Classification methods so far

New topics:

- Generative modeling
- Naïve Bayes Classifier

Alternative viewpoint

- Another way to model the inference system is to make the following assumption: $P(X, Y) = P(Y \mid X)P(X)$
- There is some underlying joint distribution p(x,y) that jointly specifies x and y.
- Write it as: $P(X, Y) = P(X \mid Y)P(Y)$
- Recall the Bayes rule: $P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$

Generative modeling

- Advantage: can generate new data points (as many as you want) based on this model

- Disadvantage: once we make a choice what distribution is data from (e.g.,Bernoulli, Gaussian, etc.), that assumption stays forever.

Generative modeling

- Since we have a probabilistic modeling, we will typically look to maximize the following:

N data points (x_i, y_i), $y_i \in \{0, 1\}$

$$\prod_{i=1}^{N} P(x_i, y_i)$$

assume data pts are $\perp$

- This is known as maximizing the joint likelihood.
- When more convenient, we will maximize the log of the above expression. $\log\prod \rightarrow \sum\log$
- Maximization done by taking derivatives.

Today's Lecture

Recap:

- Classification methods so far

New topics:

- Generative modeling
- Naïve Bayes Classifier

Naïve Bayes Classifier Applications

- Spam filtering
- Text classification
- Medical diagnosis and prediction
- Rating system

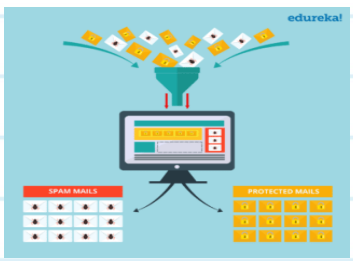
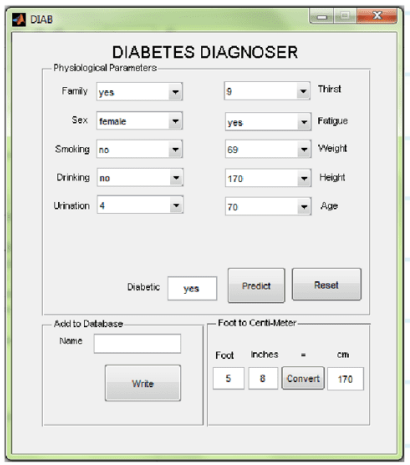


Fig. 1: MATLAB Program in execution



6. CONCLUSION

https://medium.datadriveninvestor.com/understanding-naive-bayes-and-its-application-in-text-classification-99c38e739f88

https://dzone.com/articles/naive-bayes-tutorial-naive-bayes-classifier-in-pyt

https://www.semanticscholar.org/paper/Intelligent-Na%C3%AFve-Bayes-Approach-to-Diagnose-Type-2-Sarwar-Sharma/2666863951eff26307807532315449f8b78bb4f5

Naïve Bayes Classifier

- The key assumption is **conditional independence**.
- Math: $x \in \mathbb{R}^d$

  $y \in \{0, 1\}$ label of x

  $x_i \perp x_j \mid Y$ $i \neq j$, $1 \leq i, j \leq d$

Naïve Bayes Classifier

- The key assumption is **conditional independence**.

Practical example: bag of words

- Email spam filtering based on the words present in the email.
- Suppose x_1, x_2,…,x_d are indicators of d dictionary words, so that x_i is 1 if the i-th dictionary word is present, and x_i is 0 if absent.
- Naïve Bayes assumption: $x_i \perp x_j \mid Y$

$x_i \in \{0, 1\}$

$1 \leq i \leq d$

```
presence / absence of word "vacation" is
conditionally independent from the word
"school" given that the email is
classified as say "non-spam"
```

A quick example of conditional independence

consider RVs : X, Y, Z

$$P(X, Y, Z) = P(X)P(Y|X)P(Z|Y, X)$$

if $Z \perp Y | X$:

$$P(X, Y, Z) = P(X)P(Y|X)P(Z|X)$$

Conditional independence does not imply independence

$X \perp Y \mid Z$ (cond. ind.)

vs.

$X \perp Y$ (ind.)

Likelihood

- Consider the likelihood for one example (e.g., one email):

$$P(X_1, \ldots, X_d, Y) \qquad P(Y) \cdot \prod_{i=1}^{N} P(X_i \mid Y)$$

$$P(Y) \cdot P(X_1 \mid Y) \cdot P(X_2 \mid \cancel{X_1}, Y) \cdot \ldots \cdot P(X_d \mid \cancel{X_1, \ldots, X_{d-1}}, Y)$$

- Now suppose we have N examples (e.g., N emails):

$$\prod_{k=1}^{N} P(x_k, y_k) = \prod_{k=1}^{N} P(y_k) \prod_{i=1}^{N} P(x_{k,i}, y_k)$$

dim = d

Derivations for N=1 example

- Suppose we are given prior probabilities for the label, as follows:

$$y \in \{0, 1\} \qquad y \sim \text{Bernoulli}$$
$$P(Y = 1) = \theta_0$$
$$P(Y = 0) = 1 - \theta_0$$

- Convenient representation of the above:

$$P(Y) = \theta_0^{\mathbb{1}[Y=1]} \cdot (1 - \theta_0)^{\mathbb{1}[Y=0]}$$

$$\mathbb{1}[a] = \begin{cases} 1 & \text{if } a = T \\ 0 & \text{if } a = F \end{cases}$$

Derivations for N=1 example, continued

- Notation for conditional probabilities (4 total – why ?):

$$P(Y=1) = \theta_0^1 (1-\theta_0)^0 = \theta_0$$
$$P(Y=0) = \theta_0^0 (1-\theta_0)^1 = 1-\theta_0$$

$1 \le j \le d$

$$P(X_j = 1 \mid Y=0) = \theta_{j,0}$$
$$P(X_j = 0 \mid Y=0) = 1-\theta_{j,0}$$

$$P(X_j = 1 \mid Y=1) = \theta_{j,1}$$
$$P(X_j = 0 \mid Y=1) = 1-\theta_{j,1}$$

Derivations for N=1 example, continued

- What is the total number of parameters ?

θ_0 1
$\theta_{j,0}$ d
$\theta_{j,1}$ d

} $2d+1$ parameters

Derivations for N=1 example, continued

- Write the joint probability:

$$P(X, Y) = P(Y) \prod_{i=1}^{d} P(X_j | Y)$$

- Expand:

$$\underbrace{\theta_0^{\mathbb{1}[Y=1]} \cdot (1-\theta_0)^{\mathbb{1}[Y=0]}}_{P(Y)} \cdot \prod_{i=1}^{d} \theta_{j,Y}^{\mathbb{1}[X_j=1]} \cdot (1-\theta_{j,Y})^{\mathbb{1}[X_j=0]}$$

- Note that this is a compact representation of the joint probability: one of the two product terms will hold, but not both simultaneously.

Ex. $d = 2$ $\quad (x_1, x_2, y) = (1, 0, 1)$

$$\theta_0 \cdot \theta_{1,1} \cdot (1-\theta_{2,1})$$

Derivations for N=1 example, continued

- If we take the log, we get:

$$\mathbb{1}[Y=1]\log\theta_0 + \mathbb{1}[Y=0]\log(1-\theta_0) + \sum_{j=1}^{d}\left(\mathbb{1}[X_j=1]\log\theta_{j,Y} + \mathbb{1}[X_j=0]\log(1-\theta_{j,Y})\right)$$

- This will also serve as an auxiliary result for what we will do next.

$$\prod_{i=1}^{N} \left[\theta_0^{\mathbb{1}[Y_i=1]} \cdot (1-\theta_0)^{\mathbb{1}[Y_i=0]} \cdot \prod_{j=1}^{d} \theta_{j,Y}^{\mathbb{1}[X_j=1]} \cdot (1-\theta_{j,Y})^{\mathbb{1}[X_j=0]} \right]$$

Let's now consider the case of N examples

- Now, the likelihood is: $\prod_{i=1}^{N} \left[P(Y_i) \cdot \prod_{j=1}^{d} P(X_{j,i} | Y_i) \right]$

- Take the log so we get the double summation as follows:

$$\mathbb{1}[Y_i=1] \log\theta_0 + \mathbb{1}[Y_i=0] \log(1-\theta_0)$$
$$+ \sum_{j=1}^{d} \left(\mathbb{1}[X_j=1, Y=1] \log\theta_{j,1} + \mathbb{1}[X_j=0, Y=1] \log(1-\theta_{j,1}) \right)$$
$$+ \sum_{j=1}^{d} \left(\mathbb{1}[X_j=1, Y=0] \log\theta_{j,0} + \mathbb{1}[X_j=0, Y=0] \log(1-\theta_{j,0}) \right)$$

Not as complicated as it looks!

- Useful notation:

$$N_1 = \sum_{i=1}^{N} \mathbb{1}[y_i = 1]$$

$$N_2 = \sum_{i=1}^{N} \mathbb{1}[y_i = 0]$$

$$N_1 + N_2 = N$$

recall: our objective is to estimate 2d+1 Θ parameters by maximizing joint distribution (or log of joint distribution)

Estimation of the parameter θ_0

- We are taking the derivative of the log likelihood to find the optimal choice.
- Notice that only the initial terms have θ_0 so only these terms matter when we take the derivative.

- So we can isolate:

$$\sum_{i=1}^{N} \left(1[Y_i = 1] \log \theta_0 + 1[Y_i = 0] \log(1-\theta_0) \right.$$
$$\left. + \text{other terms that do not depend on } \theta \right)$$

$$N_1 \log \theta_0 + N_2 \log(1-\theta_0) + \text{other terms}$$

Estimation of the parameter θ_0

- When we take the derivative, we get:

$$N_1 \cdot \frac{1}{\theta_0} + N_2 \cdot \frac{-1}{1-\theta_0} = 0$$

$$\theta_0(N_1 + N_2) = 1$$

$$\theta_0 = \frac{N_1}{N}$$

- This makes sense!

Estimation of the parameter $\theta_{1,1}$

- Let's now consider parameter $\theta_{1,1}$:

$$1 - \theta_{1,1} = \frac{P(X_1 = 1 | Y = 1)}{P(X_1 = 0 | Y = 1)}$$

Estimation of the parameter $\theta_{1,1}$

- Again, can isolate the terms in the log likelihood that matter for the derivative:

$$\sum_{i=1}^{N} (\text{other stuff} + 1[X_{1,i}=1, Y=1] \log\theta_{1,1} + 1[X_{1,i}=0, Y=1] \log(1-\theta_{1,1}))$$

Estimation of the parameter $\theta_{1,1}$

- Useful notation:

$$N_1 = \sum_{i=1}^{N} 1[X_{1,i} = 1 \mid Y_i = 1]$$

$$N_0 = \sum_{i=1}^{N} 1[X_{1,i} = 0 \mid Y_i = 1]$$

- Taking the derivative yields:

$$N_1 + N_0 \leq N$$

relevant term yields:

$$N_1 \log \theta_{1,1} + N_0 \log(1 - \theta_{1,1})$$

$$\Rightarrow \quad \theta_{1,1} = \frac{N_1}{N_1 + N_0} = \frac{\sum 1[X_{1,i} = 1 \mid Y_i = 1]}{\sum 1[Y_i = 1]}$$

Estimation of the remaining parameters

- Analogous to the previous case:

observe that the log likelihood express is symmetric in θ_{j, 1} / θ_{j, 0} arguments

$$\hat{\theta}_{j,1} = \frac{\sum \mathbb{1}[X_{j,i}=1 \mid Y_i=1]}{\sum \mathbb{1}[Y_i=1]}$$

$$\hat{\theta}_{j,0} = \frac{\sum \mathbb{1}[X_{j,i}=1 \mid Y_i=0]}{\sum \mathbb{1}[Y_i=0]}$$

Classification rule at testing time

- At test time, we have the following rule:

$$P(Y=1 \mid X_{test}) \underset{y=0}{\overset{y=+1}{\gtrless}} P(Y=0 \mid X_{test})$$

$$\frac{P(X_{test} \mid Y=1)P(Y=1)}{P(X_{test})} \underset{y=0}{\overset{y=+1}{\gtrless}} \frac{P(X_{test} \mid Y=0)P(Y=0)}{P(X_{test})}$$

Classification rule at testing time

- At test time, we have the following rule:

$$P(X_{test} \mid Y=1) \cdot \hat{\theta}_0 \underset{y=0}{\overset{y=1}{\gtrless}} P(X_{test} \mid Y=0) \cdot (1-\hat{\theta}_0)$$

$$P(X_{test} \mid Y=1) = \prod_{j=1}^{d} P(X_j \mid Y=1)$$

$$= \prod_{j=1}^{d} \hat{\theta}_{j,1}^{\mathbb{1}[X_j=1]} \cdot (1-\hat{\theta}_{j,1})^{\mathbb{1}[X_j=0]}$$

same for $P(X_{test} \mid Y=0)$

Description of the decision boundary

we get the equality $P(Y=1)\prod_{j=1}^{d} P(X_j|Y=1) = P(Y=0)\prod_{j=1}^{d} P(X_j|Y=0)$

apply log to both sides:

$$\log\left[\theta_0 \prod_{j=1}^{d} \theta_{j,1}^{\mathbb{I}[X_j=1]} \cdot (1-\theta_{j,1})^{\mathbb{I}[X_j=0]}\right] - \log\left[(1-\theta_0)\prod_{j=1}^{d} \theta_{j,0}^{\mathbb{I}[X_j=1]} \cdot (1-\theta_{j,0})^{\mathbb{I}[X_j=0]}\right] = 0$$

(algebra) ⋮

$$\log\frac{\theta_0}{1-\theta_0} + \sum_{i=1}^{d} \mathbb{I}[X_j=1]\log\frac{\theta_{i,1}}{\theta_{i,0}} + \sum_{i=1}^{d} \mathbb{I}[X_j=0]\log\frac{1-\theta_{i,1}}{1-\theta_{i,0}} = 0$$

$$c + \sum_{i=1}^{d} x_i v_1 + \sum_{i=1}^{d} (1-x_i) v_2 = 0$$

write as $w^T x + b \gtrless 0$ linear decision boundary

map $c, v_1, v_2,$ to w, b

Laplace smoothing

- What if there was no instance of $x_j = 1$ in the training set in either class ?
- Recall math:

$x_j = +1 \quad y = 1$

$x_j = +1 \quad y = 0$

in the email spam example, some words didn't appear in either class

$\hat{\theta}_{j,1} = 0 \qquad \hat{\theta}_{j,0} = 0$

Laplace smoothing

- Force the estimates to not be strictly zero (or one).

- Example: $N = 20$

Suppose we start with

$$\sum_{i=1}^{N} 1[Y_i = 0] = 12$$

$$\sum_{i=1}^{N} 1[Y_i = 1] = 8$$

$$\sum_{i=1}^{N} 1[X_{j,i} = 0 \mid Y_i = 1] = 0, \quad \sum_{i=1}^{N} 1[X_{j,i} = 1 \mid Y_i = 1] = 0$$

$$\hat{\theta}_{j,1} = \frac{0 + 1}{12 + 2} = 0 \qquad \hat{\theta}_{j,0} = \frac{0 + 1}{8 + 2} = 0$$

$$\frac{0}{12} \to \frac{1}{14} \qquad \frac{0}{8} \to \frac{1}{10}$$

Summary

- We introduced generative modeling.
- We discussed the special case of Naïve Bayes classifier with binary features which are conditionally independent given the class label.
- Next, class conditional densities will be Gaussian.

ECE M146 Introduction to Machine Learning

Lecture 12 - Spring 2021

Prof. Lara Dolecek

ECE Department, UCLA

Today's Lecture

Recap:

- Naïve Bayes Classifier

New topic:

- Gaussian Discriminant Analysis

Naïve Bayes Classifier

- Example of probabilistic generative modeling

$$P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

- Uses conditional independence

$$X_i \perp X_j \mid Y$$

different attributes X_i, X_j are conditionally independent given Y

- Estimate the parameters from these distributions by maximizing log of the joint distribution.

- This is done by taking derivatives (scalars, so in the usual sense).

Naïve Bayes Classifier

- Bernoulli RV models p(y) – binary classification

$$\theta_0 = P(Y=1) \qquad 1-\theta_0 = P(Y=0)$$

- Bernoulli RV models conditional probability $p(x_j=1 \mid y)$

$$P(X_j=1 \mid Y=0) = 1-\theta_{j,0} \qquad P(X_j=1 \mid Y=1) = \theta_{j,1}$$

$$P(X_j=0 \mid Y=0) = 1-\theta_{j,0} \qquad P(X_j=0 \mid Y=1) = \theta_{j,1}$$

Naïve Bayes Classifier

- End result are estimates that correspond to sample frequency.

Ex. $\hat{\theta}_0 = \frac{N_1}{N}$ ↗ # pts with Y=1 ↳ total # pts

Today's Lecture

Recap:

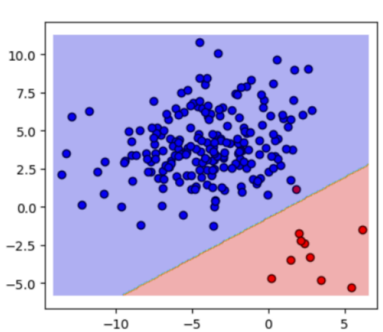
- Naïve Bayes Classifier

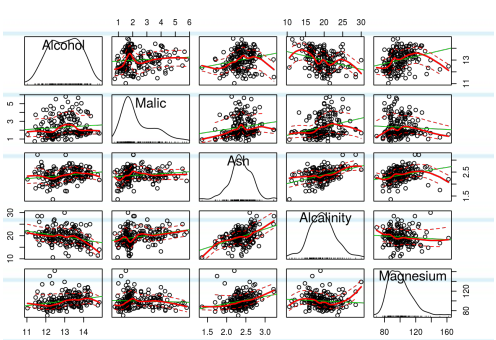
New topic:

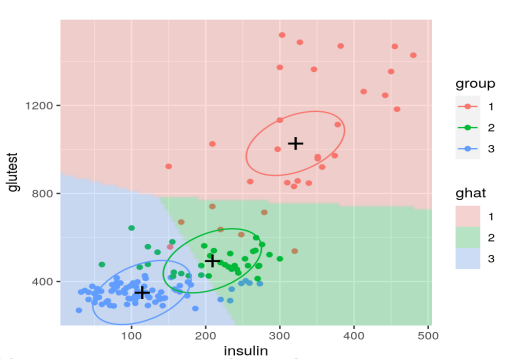
- Gaussian Discriminant Analysis

Applications abound

- Wine categorization (multivariate chemical substances are input, cultivar as class)
- Medical diagnosis of diabetes (glucose and insulin levels are input, diagnosis as class)







https://rstudio-pubs-static.s3.amazonaws.com/35817_2552e05f1d4e4db8ba87b334101a43da.html

https://bioconductor.org/help/course-materials/2017/CSAMA/labs/5-friday/lab-07-machine-learning/lab7-machine-learning.pdf

Gaussian generative modeling

- Used for binary and multiclass classification.
- Assumption is that the data in the given class is derived from a Gaussian distribution.
- Let's consider binary classification first:

$$P(X, Y) = P(Y)\,P(X \mid Y)$$

$P(Y)$: Bernoulli(θ)

$P(X \mid Y)$: Gaussian

$\mathcal{N}(\mu_1, \sigma_1^2)$

$\mathcal{N}(\mu_0, \sigma_0^2)$

Recall Gaussian RV

continuous RV

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$E(X) = \mu$$

$$\text{var}(X) = \sigma^2$$

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

Gaussian generative modeling

- Expression for the joint distribution:

suppose we have N points,
of which M are in class 1

$$f((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)) = \prod_{i=1}^{N} P(X_i \mid Y_i) P(Y_i)$$

$P(X_i \mid Y_i)$: $\mathcal{N}(\mu_1, \sigma_1^2)$ or $\mathcal{N}(\mu_0, \sigma_0^2)$

$P(Y_i)$: $\text{Bern}(\theta)$

Visualization

- Class label of y is in {0,1}, where p(y=1) = θ
- Conditional probability

M examples in class 1

$$f(x_1, x_2, \ldots x_M \mid y=1) = \prod_{i=1}^{M} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(x_i-\mu_i)^2}{2\sigma_i^2}}$$

we need to estimate μ_1 and σ_1^2 based on these M data points

Fitting a Gaussian distribution for each class

- Suppose there are M points in class y=1.
- Write the expression for the conditional density.

$$f(x_1, x_2, \ldots x_M \mid y=1) = \prod_{i=1}^{M} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(x_i-\mu_i)^2}{2\sigma_i^2}}$$

- How to find the best estimate ?

take derivative

Sample mean

$$\frac{\partial f}{\partial \mu_1} = 0$$

$$= \frac{\partial}{\partial \mu_1}\left(\prod_{i=1}^{M} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x_i-\mu_1)^2}{2\sigma_i^2}}\right)$$

take the log to bring the exponent down (this result is the same as what we would have obtained by taking the derivative of the original expression without log, because log is monotonic)

Sample mean, continued

$$\frac{\partial}{\partial \mu_1}\left(\sum_{i=1}^{M}\left(\log\frac{1}{\sqrt{2\pi\sigma_i^2}}+\log\left(e^{\frac{-(x_i-\mu_1)^2}{2\sigma_i^2}}\right)\right)\right)=0$$

$$\frac{\partial}{\partial \mu_1}\left(\sum_{i=1}^{M}\left(\log\frac{1}{\sqrt{2\pi\sigma_i^2}}-\frac{(x_i-\mu_1)^2}{2\sigma_i^2}\right)\right)=0$$

$$\frac{\partial}{\partial \mu_1}\left(-\sum_{i=1}^{M}\frac{(x_i-\mu_1)^2}{2\sigma_i^2}\right)=0$$

$$\sum_{i=1}^{M}\frac{x_i-\mu_1}{\sigma_i^2}=0$$

$$\sum_{i=1}^{M}(x_i-\mu_1)=0$$

$$\hat{\mu}_1=\frac{1}{M}\sum_{i=1}^{M}x_i$$

Sample variance

$$\frac{\partial}{\partial \sigma_1}\left(\sum_{i=1}^{M}\left(\log\frac{1}{\sqrt{2\pi\sigma_1^2}}+\log\left(e^{\frac{-(x_i-\mu_1)^2}{2\sigma_1^2}}\right)\right)\right)=0$$

$$-\frac{\partial}{\partial \sigma_1}\left(\sum_{i=1}^{M}\left(\log\sqrt{2\pi\sigma_1^2}+\frac{(x_i-\mu_1)^2}{2\sigma_1^2}\right)\right)=0$$

$$\sum_{i=1}^{M}\left(\frac{1}{\sigma_1}+\frac{(x_i-\mu_1)^2}{\sigma_1^3}\right)=0$$

$$\sum_{i=1}^{M}\frac{\sigma_1^2+(x_i-\mu_1)^2}{\sigma_1^3}=0$$

$$\sum_{i=1}^{M}\left(\sigma_1^2+(x_i-\mu_1)^2\right)=0$$

$$\hat{A}=\sigma_1^2=\frac{1}{M}\sum_{i=1}^{M}(x_i-\hat{\mu}_1)^2$$

use $\hat{\mu}_1$ if μ_1 is not available

Class parameter estimate

- Same as what we did last time for Naïve Bayes (with Bernoulli conditionals) – why ?

so far: $\hat{\mu}_i$, $\hat{\sigma}_i^2$

$$P(Y) = \theta(1-\theta)$$

$$P(X|Y)P(Y) = P(X,Y)$$

Parameter estimates

- Note that at this point we have all 5 parameter estimates

we know how to compute

$\underbrace{\hat{\mu}_1, \hat{\sigma}_1^2}_{\hat{\theta}}$ $\hat{\mu}_0, \hat{\sigma}_0^2$

At test time

- Compare the posteriors

$$\frac{P(Y=1 \mid X_{TEST})}{P(Y=0 \mid X_{TEST})} \underset{\text{class 0}}{\overset{\text{class 1}}{\gtrless}} 1$$

$$P(Y=1|X_{TEST}) = 1 - P(Y=0|X_{TEST})$$

At test time, continued

$$\frac{P(Y=1|X_T)}{P(Y=0|X_T)} = \frac{P(Y=1)P(X_T|Y=1)}{P(Y=0)P(X_T|Y=0)}$$

$$\frac{\theta}{1-\theta} \cdot \frac{\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(X_T-\mu_1)^2}{2\sigma_1^2}}}{\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(X_T-\mu_0)^2}{2\sigma_0^2}}} \gtrless 1$$

class 1 (above $\gtrless$), class 0 (below $\gtrless$)

$$\frac{\theta}{1-\theta} \cdot \frac{e^{-\frac{(X_T-\mu_1)^2}{2\sigma_1^2}}}{e^{-\frac{(X_T-\mu_0)^2}{2\sigma_0^2}}} \gtrless 1$$

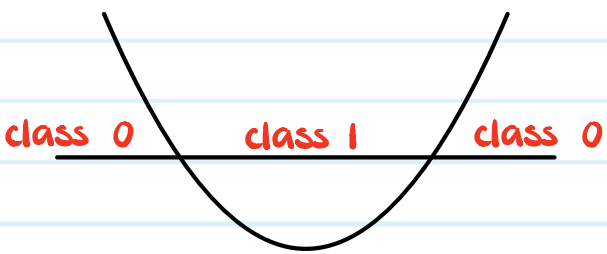
class 1 (above $\gtrless$), class 0 (below $\gtrless$)

At test time, continued

$$\underbrace{\log \frac{\theta}{1-\theta} \cdot \frac{\sigma_0}{\sigma_1}}_{=-\tau} + \frac{(x_T-\mu_0)^2}{2\sigma_0^2} - \frac{(x_T-\mu_1)^2}{2\sigma_1^2} \underset{\text{class 0}}{\overset{\text{class 1}}{\gtrless}} 0$$

$$\frac{(x_T-\mu_0)^2}{2\sigma_0^2} - \frac{(x_T-\mu_1)^2}{2\sigma_1^2} \underset{\text{class 0}}{\overset{\text{class 1}}{\lessgtr}} \tau$$

quadratic w.r.t. x_T



Now, let's consider the case of both classes having the same variance

- Write the joint pdf:

$$\prod_{i=1}^{N} P(x_i, y_i) = \prod_{i=1}^{N} \theta^{1[y_i=1]} \cdot (1-\theta)^{1[y_i=0]} \cdot \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu_1)^2}{2\sigma^2}}\right)^{1[y_i=1]} \cdot \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu_0)^2}{2\sigma^2}}\right)^{1[y_i=0]}$$

- How many parameters to be estimated do we have now ? 4 : $\theta, \mu_0, \mu_1, \sigma^2$
- Take the log, as before.

Analysis continued

$$\sum_{i=1}^{N} \Big(\mathbb{1}[y_i = 1] \log \theta + \mathbb{1}[y_i = 0] \log(1-\theta) + \mathbb{1}[y_i = 1] \log\Big(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu_1)^2}{2\sigma^2}}\Big) + \mathbb{1}[y_i = 0] \log\Big(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu_0)^2}{2\sigma^2}}\Big) \Big)$$

take the derivative w.r.t. each parameter and set to 0

$$\hat{\theta} = \frac{M}{N}$$

$$\hat{\mu}_1 = \frac{1}{M} \sum_{i=1}^{N} x_i \cdot \mathbb{1}[Y_i = 1]$$

$$\hat{\mu}_0 = \frac{1}{M} \sum_{i=1}^{N} x_i \cdot \mathbb{1}[Y_i = 0]$$

Analysis continued

- Estimate of the variance is new: let $t = \sigma^2$

$$\sum_{i=1}^{N} \Big(1[y_i = 1] \log\theta + 1[y_i = 0] \log(1-\theta) + 1[y_i = 1] \log\Big(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu_1)^2}{2\sigma^2}}\Big) + 1[y_i = 0] \log\Big(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu_0)^2}{2\sigma^2}}\Big)\Big)$$

$$\sum_{i=1}^{N} \Big(1[y_i = 1]\Big(-\frac{1}{2}\log(2\pi t) - \frac{(x_i-\mu_1)^2}{2t}\Big) + 1[y_i = 0]\Big(-\frac{1}{2}\log(2\pi t) - \frac{(x_i-\mu_0)^2}{2t}\Big)\Big)$$

relevant terms only

Analysis continued

$$\sum_{i=1}^{N}\left(1[y_i=1]\left(\frac{-1}{2t}+\frac{(x_i-\mu_1)^2}{2t^2}\right)+1[y_i=0]\left(\frac{-1}{2t}+\frac{(x_i-\mu_0)^2}{2t^2}\right)\right)=0$$

$$\sum_{i=1}^{N}\left(1[y_i=1]\left(-t+(x_i-\mu_1)^2\right)+1[y_i=0]\left(-t+(x_i-\mu_0)^2\right)\right)=0$$

$$-Nt+\sum_{i=1}^{N}\left(1[y_i=1](x_i-\mu_1)^2+1[y_i=0](x_i-\mu_0)^2\right)=0$$

$$\hat{A}=\hat{\sigma}^2=\frac{1}{N}\left(\sum_{i=1}^{N}1[y_i=1](x_i-\mu_1)^2+\sum_{i=1}^{N}1[y_i=0](x_i-\mu_0)^2\right)$$

Analysis continued

$$M = \sum_{i=1}^{N} 1[y_i = 1]$$

$$N - M = \sum_{i=1}^{N} 1[y_i = 0]$$

$$- s_1 = \frac{1}{M} \sum_{i=1}^{N} 1[y_i = 1](x_i - \mu_1)^2$$

$$- s_0 = \frac{1}{N-M} \sum_{i=1}^{N} 1[y_i = 0](x_i - \mu_0)^2$$

$$\hat{\sigma}^2 = s_1 \frac{M}{N} + s_0 \frac{N-M}{N}$$

At test time

- Again, compare conditional probabilities:

$$P(Y=1|X) \quad \text{vs.} \quad P(Y=0|X)$$

- This is equivalent to:

$$\frac{P(Y=1|X)}{P(Y=0|X)} \propto \frac{P(Y=1,X)}{P(Y=0,X)} \gtrless 1$$

class 1

class 0

At test time, continued

$$\frac{\theta \cancel{\frac{1}{\sqrt{2\pi\sigma^2}}} e^{\frac{-(x_T-\mu_1)^2}{2\sigma^2}}}{(1-\theta)\cancel{\frac{1}{\sqrt{2\pi\sigma^2}}} e^{\frac{-(x_T-\mu_0)^2}{2\sigma^2}}} \gtrless 1$$

class 1 / class 0

$$\log\frac{\theta}{1-\theta} + \frac{(x_T-\mu_0)^2}{2\sigma^2} - \frac{(x_T-\mu_1)^2}{2\sigma^2} \gtrless 0$$

class 1 / class 0

$$\log\frac{\theta}{1-\theta} + x_T\frac{\mu_1-\mu_0}{\sigma^2} + \frac{\mu_0^2-\mu_1^2}{2\sigma^2} \gtrless 0$$

class 1 / class 0

$$x \cdot w_1 + w_0 \gtrless 0$$

class 1 / class 0

now the decision boundary is linear

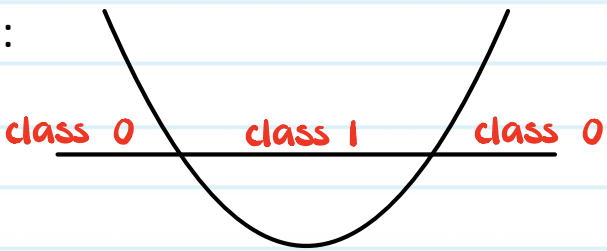
- Special case of uniform prior.

$$\theta = 0.5$$

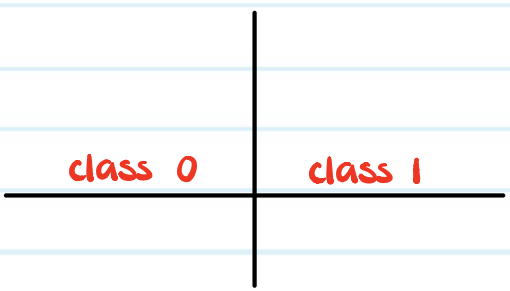
$$\log\frac{\theta}{1-\theta} = 0$$

Decision boundary

- In the case of unequal class variances:



- In the case of equal class variances:



Recap

- We studied another instance of probabilistic generative modeling, this time with in-class Gaussians.
- We saw the scalar case (single attribute). The decision boundary is either linear or quadratic depending on whether the class variances are the same or different.
- Binary classification extends to multiclass classification.

Multiclass Classification

$y \in \{0, 1, 2\}$ (no longer Bernoulli)

let's consider k = 3 classes with equal variances

$$P(X, Y) = \prod_{i=1}^{N} \theta_0 \cdot \theta_1 \cdot (1 - \theta_0 - \theta_1)$$

$\theta_0 = P(Y = 0)$ $\quad \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu_0)^2}{2\sigma^2}}$ $\quad 1[y_i = 0]$

$\theta_1 = P(Y = 1)$ $\quad \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu_1)^2}{2\sigma^2}}$ $\quad 1[y_i = 1]$

$1 - \theta_0 - \theta_1 = P(Y = 2)$ $\quad \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu_2)^2}{2\sigma^2}}$ $\quad 1[y_i = 2]$

Multiclass Classification

how many parameters do we need to estimate now?

$\theta_0, \theta_1, \mu_0, \mu_1, \mu_2, \sigma^2$

$$\hat{\theta}_0 = \frac{1}{N}\sum_{i=1}^{N} 1[y_i = 0]$$

$$\hat{\theta}_1 = \frac{1}{N}\sum_{i=1}^{N} 1[y_i = 1]$$

Multiclass Classification

equivalent to the binary case
we already studied

$$N_0 = \sum_{i=1}^{N} \mathbb{1}[y_i = 0]$$

$$N_1 = \sum_{i=1}^{N} \mathbb{1}[y_i = 1]$$

$$\hat{\mu}_0 = \frac{1}{N_0} \sum_{i=1}^{N} x_i \, \mathbb{1}[y_i = 0]$$

same for $\hat{\mu}_1, \hat{\mu}_2$

$$\hat{\sigma}^2 = s_0 \frac{N_0}{N} + s_1 \frac{N_1}{N} + s_2 \frac{N - N_0 - N_1}{N}$$

s_0, s_1, s_2 are in-class sample variances

At test time

$P(Y=0|X_T)$ vs. $P(Y=1|X_T)$ vs. $P(Y=2|X_T)$

$$f_0 = \theta_0 \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu_0)^2}{2\sigma^2}}$$

$$f_1 = \theta_1 \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu_1)^2}{2\sigma^2}}$$

$$f_2 = (1-\theta_0-\theta_1) \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu_2)^2}{2\sigma^2}}$$

At test time

$$\frac{f_0}{f_1} \gtrless 1 \qquad \frac{f_1}{f_2} \gtrless 1 \qquad \frac{f_0}{f_2} \gtrless 1$$

if for example, f_0 / f_1 > 1, we don't even need to check f_1 / f_2. check f_0 / f_2.

Recap

- Generative modeling with Gaussian class conditionals

- First: single feature `x is a scalar`
- Binary classification
- Multiclass classification

- Next: multiple features `x is a vector`