Classification (2)

Never Stand Still

COMP9417 Machine Learning & Data Mining

Term 1, 2020

Adapted from slides by Dr Michael Bain

Aims

This lecture will continue your exposure to machine learning approaches to the problem of classification. Following it you should be able to reproduce theoretical results, outline algorithmic techniques and describe practical applications for the topics:

- explain the concept of inductive bias in machine learning
- outline Bayes Theorem as applied in machine learning
- define MAP and ML inference using Bayes theorem
- define the Bayes optimal classification rule in terms of MAP inference
- outline the Naive Bayes classification algorithm
- describe typical applications of Naive Bayes for text classification
- outline the logistic regression classification algorithm

Introduction

What do we understand about the problem of learning classifiers ? . . .

how can we know when classifier learning succeeds ?

and . . . can we use this to build practical algorithms ?

Inductive Bias

"All models are wrong, but some models are useful."

Box & Draper (1987)

Inductive Bias

Confusingly, “inductive bias” is NOT the same “bias” as in the “bias-variance” decomposition.

“Inductive bias” is the combination of assumptions and restrictions placed on the models and algorithms used to solve a learning problem.

Essentially it means that the algorithm and model combination you are using to solve the learning problem is appropriate for the task.

Success in machine learning requires understanding the inductive bias of algorithms and models, and choosing them appropriately for the task.

Inductive Bias

Unfortunately, for most machine learning algorithms it is not always easy to know what their inductive bias is.

For example, what is the inductive bias of:

- Linear Regression ?
  - ○ …
  - ○ …
- Nearest Neighbour ?
  - ○ …
  - ○ …

Inductive Bias

Unfortunately, for most machine learning algorithms it is not always easy to know what their inductive bias is.

For example, what is the inductive bias of:

- Linear Regression ?
    - target function has the form $y = ax + b$
    - approximate by fitting using MSE
- Nearest Neighbour ?
    - target function is a complex non-linear function of the data predict using nearest neighbour by Euclidean distance in feature space

Inductive Bias

What we would really like:

- a framework for machine learning algorithms
- with a way of representing the inductive bias
- ideally, should be a declarative specification
- also should quantify uncertainty in the inductive bias

A probabilistic approach

A simple probabilistic model

‘Viagra’ and ‘lottery’ are two Boolean features; Y is the class variable, with values ‘spam’ and ‘ham’. In each row the most likely class is indicated in bold.

Viagra	lottery	$P(Y = \text{spam} \mid \text{Viagra}, \text{lottery})$	$P(Y = \text{ham} \mid \text{Viagra}, \text{lottery})$
0	0	0.31	**0.69**
0	1	**0.65**	0.35
1	0	**0.80**	0.20
1	1	0.40	**0.60**

Decision rule

Assuming that X and Y are the only variables we know and care about, the posterior distribution $P(Y\ |X)$ helps us to answer many questions of interest.

- For instance, to classify a new e-mail we determine whether the words 'Viagra' and 'lottery' occur in it, look up the corresponding probability $P\ (Y\ = spam|Viagra, lottery)$, and predict spam if this probability exceeds 0.5 and ham otherwise.
- Such a recipe to predict a value of Y on the basis of the values of X and the posterior distribution $P(Y\ |X)$ is called a decision rule.

Bayesian Machine Learning

Two Roles for Bayesian Methods

Provides practical learning algorithms:

- Naive Bayes classifier learning
- Bayesian network learning, etc.
- Combines prior knowledge (prior probabilities) with observed data

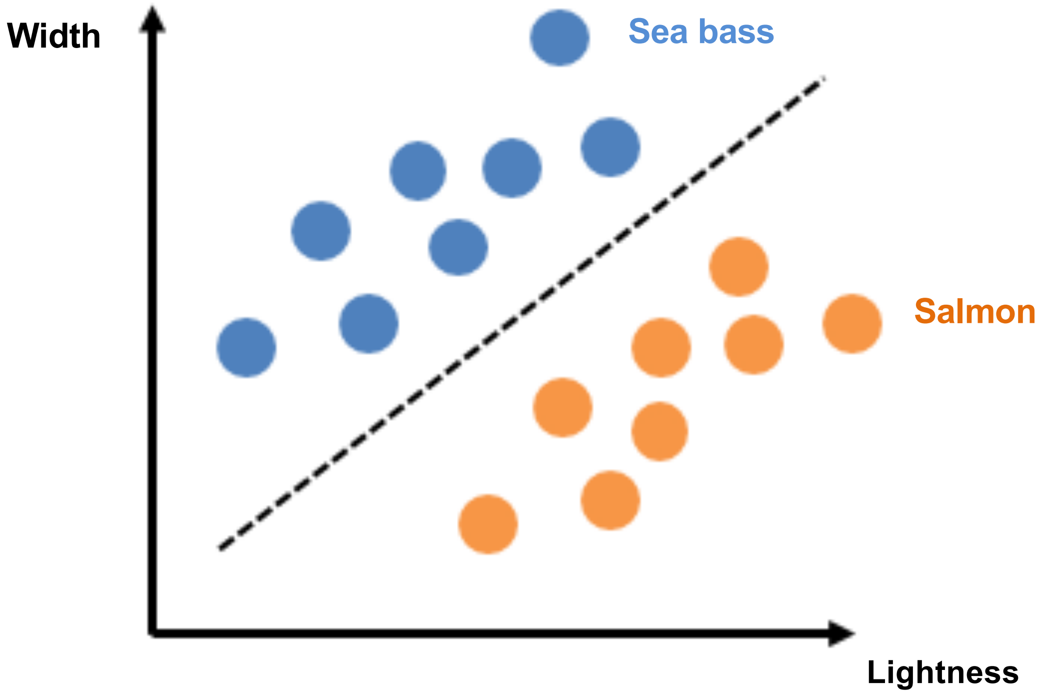
Provides useful conceptual framework:

- Provides a “gold standard” for evaluating other learning algorithms
- Some additional insight into Occam’s razor

Classification

Question:

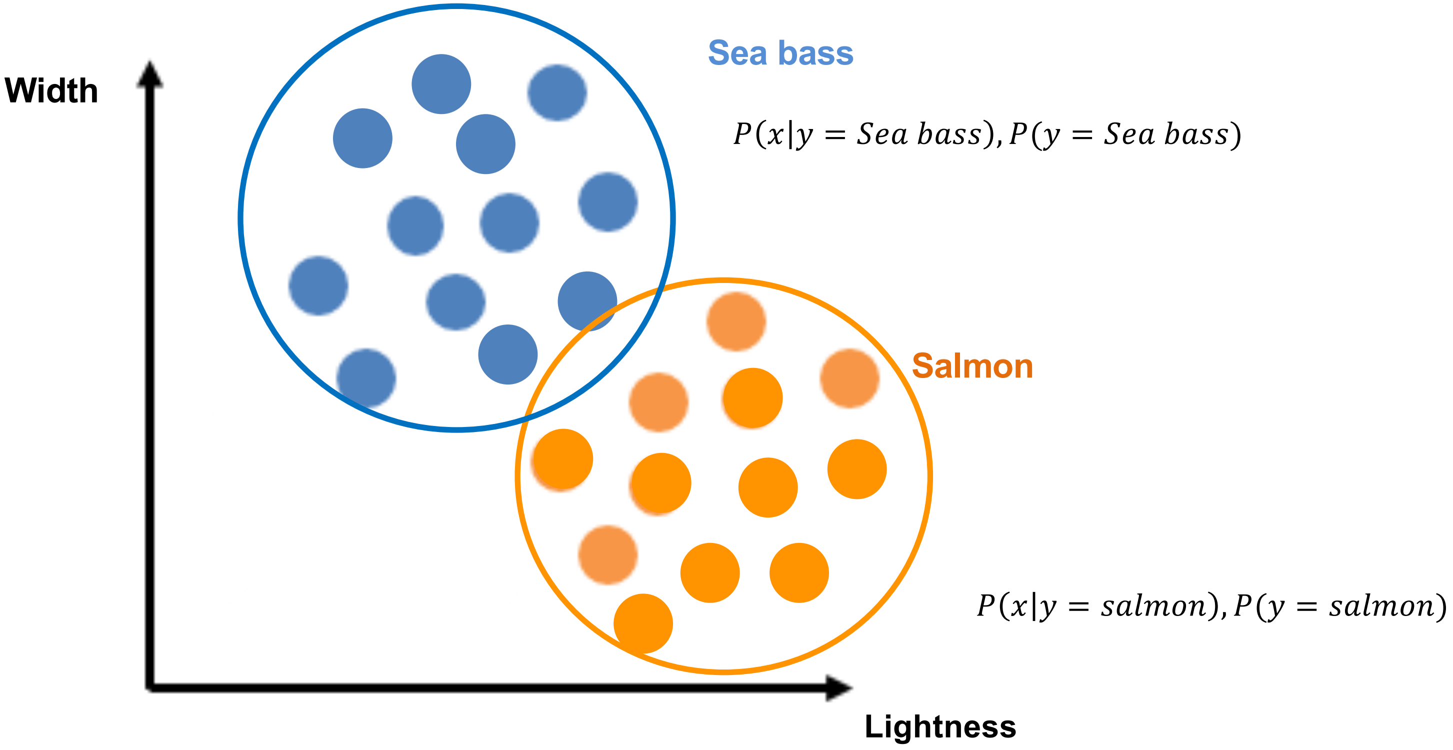
Can we do something different than finding the discriminative line (or some boundary) to be able to separate the two groups?



Classification

Answer:

Yes, we can. Instead of finding a discriminative line, maybe we can focus on one class at a time and build a model that describes how that class looks like; and then do the same for the other class. This type of models are called *generative learning algorithm*.


Sea bass
$P(x|y = Sea\ bass), P(y = Sea\ bass)$
Width
Salmon
$P(x|y = salmon), P(y = salmon)$
Lightness

Classification – Bayesian methods

Example: Imagine, we want to classify fish type: Salmon , Sea bass

If from the past experience we have (C_i $is\ the\ class$):

P(c_i)	Salmon	Sea bass
Prior	0.3	0.7

- If we decide only based on prior, we always have to choose "sea bass". This is called "*decision rule based on prior*"
  - This can behave vey poorly
  - It never predicts other classes

Classification – Bayesian methods

Example: now if we have some more information on the length of the fish in each class, then how can we update our decision, if we want to predict the class for a fish with 70cm length?

These are called “class conditionals”, “class conditioned probabilities”

$\boldsymbol{P(x\|c_i)}$	Salmon	Sea bass
length > 100 cm	0.5	0.3
50 cm < length < 100 cm	0.4	0.5
length < 50 cm	0.1	0.2

What we are interested in is $P(c_i|x)$, but we have $P(c_i)$ and $P(x|c_i)$, so how should we go about this?

Bayes Theorem

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

where:

$P(h) =$ prior probability of hypothesis h

$P(D) =$ prior probability of training data D

$P(h|D) =$ probability of h given D

$P(D|h) =$ probability of D given h

Decision Rule from Posteriors

Example:

If the output belongs to a set of k classes: $y \in \{C_1, C_2, \dots, C_k\}$ for $1 \leq i \leq k$

Then in Bayesian framework:

$$P(y = C_i|x) = \frac{P(x|C_i).P(C_i)}{P(x)}$$

- $P(y = C_i|x)$: posterior probability
- $P(x|C_i)$: class conditional (likelihood)
- $P(C_i)$: prior
- $P(x)$: Marginal ($P(x) = \sum_i p(x|C_i).P(C_i)$)

The decision rule is to select a class which maximizes the posterior probability for the prediction

Choosing Hypotheses

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Generally want the most probable hypothesis given the training data, Maximum a posteriori hypothesis h_{MAP} :

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h|D) \\ &= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D|h)P(h) \end{aligned}$$

Choosing Hypotheses

If assume $P(h_i) = P(h_j)$ then can further simplify, and choose the *Maximum likelihood* (ML) hypothesis:

$$h_{ML} = \arg\max_{h_i \in H} P(D|h_i)$$

Classification – Bayesian methods

Example: now if we have some more information on the length of the fish in each class, then how can we update our decision, if we want to predict the class for a fish with 70cm length?

$\boldsymbol{P(x \mid c_i)}$	Salmon	Sea bass
length > 100 cm	0.5	0.3
50 cm < length < 100 cm	0.4	0.5
length < 50 cm	0.1	0.2

$$P(c = salmon \mid x = 70cm) \propto P(70cm \mid salmon) * P(salmon) = 0.4*0.3=0.12$$
$$P(c = sea\ bass \mid x = 70cm) \propto P(70cm \mid sea\ bass) * P(sea\ bass) = 0.5*0.7=0.35$$

So base on these probabilities, our model predict the type as “sea bass”

Applying Bayes Theorem

Does patient have cancer or not?

A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have this cancer.

$P(cancer) = ?$ $\quad$ $P(not\ cancer) = ?$

$P(\oplus\ |cancer) = ?$ $\quad$ $P(\ominus\ |cancer) = ?$

$P(\oplus\ |not\ cancer) = ?$ $\quad$ $P(\ominus\ |not\ cancer) = ?$

Applying Bayes Theorem

Does patient have cancer or not?

A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have this cancer.

$$P(cancer) = .008 \qquad P(not\ cancer) = .992$$

$$P(\oplus\ |cancer) = .98 \qquad P(\ominus\ |cancer) = 0.02$$

$$P(\oplus\ |not\ cancer) = .03 \qquad P(\ominus\ |not\ cancer) = .97$$

Applying Bayes Theorem

Does patient have cancer or not?

$$P(cancer|\oplus) = ?$$

$$P(not\ cancer|\oplus) = ?$$

We can find the maximum a posteriori (MAP) hypothesis

$$P(\oplus|cancer)P(cancer) = 0.98 \times 0.008 = 0.00784$$

$$P(\oplus|not\ cancer)P(not\ cancer) = 0.03 \times 0.992 = 0.02976$$

Thus $h_{MAP} = \cdots$

Applying Bayes Theorem

Does patient have cancer or not?

$$P(cancer|\oplus) = ?$$

$$P(not\ cancer|\oplus) = ?$$

We can find the maximum a posteriori (MAP) hypothesis

$$P(\oplus|cancer)P(cancer) = 0.98 \times 0.008 = 0.00784$$

$$P(\oplus|not\ cancer)P(not\ cancer) = 0.03 \times 0.992 = 0.02976$$

Thus $h_{MAP} = not\ cancer$

Applying Bayes Theorem

How to get the posterior probability of a hypothesis h ?

Divide by $P\ (\bigoplus)$ (probability of data) to normalize result for h:

$$P(h|D) = \frac{P(D|h)P(h)}{\sum_{h_i \in H} P(D|h_i)P(h_i)}$$

Denominator ensures we obtain posterior probabilities that sum to 1.

Sum for all possible numerator values, since hypotheses are mutually exclusive (e.g., patient either has cancer or does not).

Basic Formulas for Probabilities

Product Rule: probability $P\ (A\ \wedge\ B)$ of conjunction of two events A and B:

$$P(A\ \wedge\ B)\ =\ P(A|B)P(B)\ =\ P(B|A)P(A)$$

Sum Rule: probability of disjunction of two events A and B:

$$P\ (A\ \vee\ B)\ =\ P\ (A)\ +\ P\ (B)\ -\ P\ (A\ \wedge\ B)$$

Theorem of total probability: if events $A_1, \ldots, A_n$ are mutually exclusive with $\sum_{i=1}^{n} P(A_i) = 1$, then:

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Basic Formulas for Probabilities

Also worth remembering:

- Conditional Probability: probability of A given B:

$$\mathrm{P(A|B)} = \frac{P(A \wedge B)}{P(B)}$$

- Rearrange sum rule to get:

$$P\,(A\ \wedge\ B)\ =\ P\,(A)\ +\ P\,(B)\ -\ P\,(A\ \vee\ B)$$

Posterior Probabilities & Decision

If we have two competing hypothesis h_1 and h_2, then

- $P(h_1|D)$ = posterior probability of h_1
- $P(h_2|D)$ = posterior probability of h_2

- So far, the potential decision is made based on the higher posterior probability
  - Decide h_1 if $P(h_1|D)$> $P(h_2|D)$ and else decide h_2
- An alternative: using a loss function $L(h)$, where $L(h)$ is the loss that occurs when decision h is made.
  - In this setup the Bayesian testing procedure minimizes the posterior expected loss

Bayesian Expected Loss

Example: If in the example of patient with positive test result, we know that the cost of misclassifying a patient who has cancer as “not cancer” is 10 times more than misclassifying a patient who doesn’t have cancer as “cancer”, how that will affect our decision?

Bayesian Expected Loss (Risk)

If the cost of misclassification is not the same for different classes, then instead of maximizing a posteriori, we have to minimize the expected loss:

- So if we define the loss associated to action α_i as $\lambda(\alpha_i|h)$
- Then the expected loss associated to action α_i is:

$$E[L(\alpha_i)] = R(\alpha_i|x) = \sum_{h \in H} \lambda(\alpha_i|h)\ P(h|x)$$

An optimal Bayesian decision strategy is to minimize the expected loss. And if the loss associated to misclassification is the same for different classes, then maximum a posteriori is equal to minimizing the expected loss.

Bayesian Expected Loss

Example: let's revisit the example of patient with positive result for cancer, given the loss function below:

$\lambda(\alpha_i\|c_i)$	Cancer	Not cancer
If predicted cancer	0	1
If predicted not cancer	10	0

$$R(predict\ cancer|\oplus) = \lambda(predict\ cancer|cancer)P(cancer|\oplus)$$
$$+\lambda(predict\ cancer|not\ cancer)P(not\ cancer|\oplus) \propto 0 + 1\times0.02976 = 0.02976$$

$$R(predict\ not\ cancer|\oplus) = \lambda(predict\ not\ cancer|cancer)P(cancer|\oplus)$$
$$+\lambda(predict\ not\ cancer|not\ cancer)P(not\ cancer|\oplus) \propto 10\times0.00784 + 0 = 0.0784$$

$$R(predict\ cancer|\oplus) < R(predict\ not\ cancer|\oplus)$$

Therefor the expected loss is less if we predict that the patient has cancer.

Bayesian Expected Loss

In summary:

- Bayesian framework allows for integration of losses into the decision-making rule
- In such framework, we minimize the posterior expected loss

Bayes Theorem

- To compute $P(D|h)$ and $P(h)$, we can use an empirical method based on given data
- Or we may assume a parametric model, then we estimate parameters using the data

Learning A Real Valued Function

Consider any real-valued target function f
Training examples $\langle x_i, y_i \rangle$, where y_i is noisy training value

- $y_i = f(x_i) + \varepsilon_i$
- ε_i is random variable (noise) drawn independently for each x_i according to some Gaussian (normal) distribution with mean zero

Then the maximum likelihood hypothesis h_{ML} is the one that minimizes the sum of squared errors:

$$h_{ML} = \arg\max_{h \in H} P(D|h) = \arg\max_{h \in H} \prod_{i=1}^{m} P(y_i|h)$$

$$= \arg\max_{h \in H} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{y_i - h(x_i)}{\sigma})^2}$$

Where $\hat{f} = h_{ML}$

Learning A Real Valued Function

Maximize natural log to give simpler expression:

$$h_{ML} = \arg\max_{h \in H} \sum_{i=1}^{m} ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2} (\frac{d_i - h(x_i)}{\sigma})^2$$

$$= \arg\max_{h \in H} \sum_{i=1}^{m} -\frac{1}{2} (\frac{d_i - h(x_i)}{\sigma})^2$$

$$= \arg\max_{h \in H} \sum_{i=1}^{m} -(d_i - h(x_i))^2$$

Equivalently, we can minimize the positive version of the expression:

$$h_{ML} = \arg \min_{h \in H} \sum_{i=1}^{m} (d_i - h(x_i))^2$$

Discriminative vs Generative Probabilistic Models

- Discriminative models model the posterior probability distribution $P\ (Y\ |X)$, where Y is the target variable and X are the features. That is, given X they return a probability distribution over Y .
- Generative models model the joint distribution $P\ (Y, X)$ of the target Y and the feature vector X. Once we have access to this joint distribution, we can derive any conditional or marginal distribution involving the same variables. In particular, since $P(X) = \sum_y P(Y = y, X)$ it follows that the posterior distribution can be obtained as$P(Y|X) = \frac{P(Y,X)}{\sum_y P(Y=y,X)}$
- Such models are called '*generative*' because we can sample from the joint distribution to obtain new data points together with their labels. Alternatively, we can use $P(Y)$ to sample a class and $P(X|Y)$ to sample an instance for that class.