Overview for Week 2

THE AUSTRALIAN NATIONAL UNIVERSITY

What is Learning Theory

Basic Idea, Classification, Regression, Novelty Detection, Applications

Statistics

Probabilities, Distributions, Bayes Rule, Measures

Useful Distributions

Normal Distribution, Laplacian Distribution, Mean, Variance

Maximum Likelihood Estimators

Parameter Adjustment, Density Estimation, Optimization Problems

Convergence

Law of Large Numbers, Curse of Dimensionality, Overfitting

Perceptron

Biological Background, Hebbian Learning, Half Spaces, Linear Functions, Perceptron Learning Rule, Convergence Theorem, Applications, Algorithms and Extensions.

Learning Theory

THE AUSTRALIAN NATIONAL UNIVERSITY

Supervised Learning

We have some empirical data (images, medical observations, market indicators, socioeconomical background of a person, texts), say $\mathbf{x}_i \in \mathcal{X}$, usually together with labels $y_i \in \mathcal{Y}$ (digits, diseases, share price, credit rating, relevance) and we want to find a function that connects $\mathbf{x}$ with y.

Cost of Misprediction

Typically, there will be a function $c(\mathbf{x}, y, f(\mathbf{x}))$ depending on $\mathbf{x}, y$ and the prediction $f(\mathbf{x})$ which tells us the loss we incur by estimating $f(\mathbf{x})$ instead of the actual value y. This may be, e.g. the number of wrong characters in an OCR application, the cost for wrong treatment, the actual loss of money on the stock exchange, the cost for a bankrupt client, or the amount of annoyance by receiving a wrong e-mail.

Unsupervised Learning

We have some data $\mathbf{x}_i$ and we want to find regularities or interesting objects in general from the data.

Classification

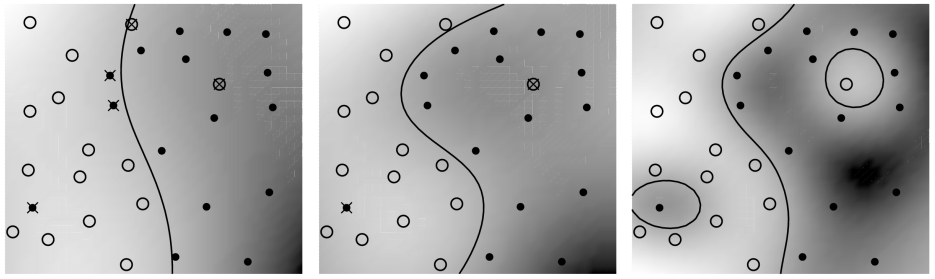
THE AUSTRALIAN NATIONAL UNIVERSITY

Goal

We want to find a function $f : \mathcal{X} \to \{\pm 1\}$ or $f : \mathcal{X} \to \mathbb{R}$ which will tell us whether a new observation $\mathbf{x} \in \mathcal{X}$ belongs to class 1 or -1 (and possibly the confidence with which this is so.

Questions

How to rate f, how to find f, how to interpret $f(\mathbf{x})$, ...



Applications

THE AUSTRALIAN NATIONAL UNIVERSITY

Optical Character Recognition

The goal is to classify (handwritten) characters (note that here f has to map into $\{a, \ldots, z\}$ rather than into $\{-1, 1\}$) automatically (form readers, scanners, post).

Spam Filtering

Determine whether an e-mail is spam or not (or whether it is urgent), based on keywords, word frequencies, special characters (\$, !, uppercase, whitespace), ...

Medical Diagnosis

Given some observations such as immune status, blood pressure, etc., determine whether a patient will develop a certain disease. Here it matters that we can estimate the probability of such an outcome.

Face Detection

Given a patch of an image, determine whether this corresponds to a face or not (useful for face tracking later — see e.g. Seeing Machines).

Learning is not (only) Memorizing

THE AUSTRALIAN NATIONAL UNIVERSITY

Goal

Infer from a training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ to the general connection between $\mathbf{x}$ and y.

Simple and Dumb Learning

Memorize all the data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ we are given. This clearly gets the hypotheses on the training set right (popular strategy in the 80s), but what to do for new $\mathbf{x}$?

Slightly Better Strategy

Memorize all the data and for new observations predict the same as the closest neighbour $\mathbf{x}_i$ to $\mathbf{x}$ does (Nearest Neighbour algorithm).

Problem

What about noisy and unreliable data? Different notions of proximity? We want to generalize our training data to new and unknown concepts.

Errors in Classification

THE AUSTRALIAN NATIONAL UNIVERSITY

Classification Error

We say that a classification error occurs, if $\operatorname{sgn} f(\mathbf{x}) \neq y$, i.e. if $\operatorname{sgn} f(\mathbf{x})$ predicts an outcome different from y.

Data Generation

Assume that the $\mathbf{x}$ are drawn from some distribution (*we will come back to that later*) and that y is either deterministic (for every $\mathbf{x}$ there is only one true y) or random ($y = 1$ occurs in a fraction p of all cases and $y = -1$ in a fraction of $1 - p$ cases).

This "distribution" could simply be due to someone writing 0 and 1 on a piece of paper. Or the different processes that smudge and degrade paper when an envelope ends up in the postbox.

Goal

We want to rate our classifier f with respect to this data generating process.

Performance Guarantees

THE AUSTRALIAN NATIONAL UNIVERSITY

Expected Error

It is the average error we can expect for new data arriving, i.e.

$$R[f] = \mathbf{E}_{(\mathbf{x},y)} \left[\operatorname{sgn} f(\mathbf{x}) \neq y\right]$$

Sometimes we also call $R[f]$ the expected risk.

Training Error

The average error on the training set

$$R_{\text{emp}}[f] := \frac{1}{m} \sum_{i=1}^{m} (\operatorname{sgn} f(x_i) \neq y_i)$$

Sometimes we also call $R_{\text{emp}}[f]$ the empirical risk.

Connection between Expected Error and Training Error

How much can we say about $R[f]$, given $R_{\text{emp}}[f]$? Clearly $R[f]$ can (and usually will) exceed the training error. We want to know the probability of this happening.

Regression

THE AUSTRALIAN NATIONAL UNIVERSITY

Goal

We want to find a function $f : \mathcal{X} \to \mathbb{R}$ which will tell us the value of f at $\mathbf{x}$.

Application: Getting Rich

Predict the stock value of IBM/CISCO/BHP/TELSTRA ... given today's market indicators (plus further background data).

Application: Wafer Fab

Predict (and optimize) the yield for a microprocessor, given the process parameters (temperature, chemicals, duration, ...).

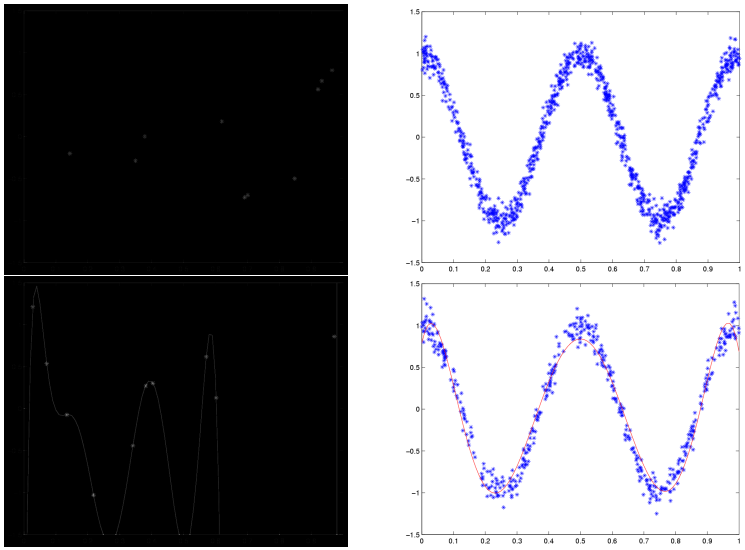
Application: Network Routers

Predict the network traffic through some hubs/routers/switches. We could reconfigure the infrastructure in time ...

Application: Drug Release

Predict the requirement for a certain drug (e.g. insulin) automatically.

Example



Novelty Detection

Goal

- Build estimator that finds unusual observations and outliers
- Build estimator that can assess how typical an observation is

Idea

- Data is generated according to some density $p(x)$. Find regions of low density.
- Such areas can be approximated as the **level set** of an auxiliary function. No need to estimate $p(x)$ directly — use proxy of $p(x)$.
- Specifically: find $f(\mathbf{x})$ such that $\mathbf{x}$ is novel if $f(\mathbf{x}) \leq c$ where c is some constant, i.e. $f(\mathbf{x})$ describes the amount of novelty.

Applications

Network Intrusion Detection

Detect whether someone is trying to hack the network, downloading tons of MP3s, or doing anything else *unusual* on the network.

Jet Engine Failure Detection

You can't (normally) destroy a couple of jet engines just to see *how* they fail.

Database Cleaning

We want to find out whether someone stored bogus information in a database (typos, etc.), mislabelled digits, ugly digits, bad photographs in an electronic album.

Fraud Detection

Credit Card Companies, Telephone Bills, Medical Records

Self calibrating alarm devices

Car alarms (adjusts itself to where the car is parked), home alarm (location of furniture, temperature, open windows, etc.)

Probability

Basic Idea

We have events, denoted by sets $X \subset \mathcal{X}$ in a space of possible outcomes $\mathcal{X}$. Then $\Pr(X)$ tells us how likely is that an event $\mathbf{x}$ with $\mathbf{x} \in X$ will occur.

Basic Axioms

- $\Pr(X) \in [0, 1]$ for all $X \subseteq \mathcal{X}$
- $\Pr(\mathcal{X}) = 1$
- $\Pr\left(\cup_i X_i\right) = \sum_i \Pr(X_i)$ if $X_i \cap X_j = \emptyset$ for all $i \neq j$

Actually, we were a bit sloppy — $\Pr(X)$ may not always defined and all we need is a σ-algebra on $\mathcal{X}$. But we won't worry about this here ...

Simple Corollary

$$\Pr(X_i \cup X_j) = \Pr(X_i) + \Pr(X_j) - \Pr(X_i \cap X_j)$$

Multiple Variables

THE AUSTRALIAN NATIONAL UNIVERSITY

Two Sets

Assume we have $\mathcal{X}$ and $\mathcal{Y}$ and a probability measure on the product space of $\mathcal{X}$ and $\mathcal{Y}$. Then we may consider the space of events $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ and ask how likely events in such a space are.

Independence

If the events $\mathbf{x}$ and $\mathbf{y}$ are independent, for any sets $X \subset \mathcal{X}$ and $Y \subset \mathcal{Y}$ we have

$$\Pr(X, Y) = \Pr(X) \cdot \Pr(Y).$$

Here $\Pr(X, Y)$ is the probability that any $(\mathbf{x}, \mathbf{y})$ with $\mathbf{x} in X$ and $\mathbf{y} \in Y$ occur.

Dependence and Conditional Probability

Typically, knowing $\mathbf{x}$ will tell us something about $\mathbf{y}$ (think regression or classification). We have

$$\Pr(Y|X)\Pr(X) = \Pr(Y, X).$$

This implies $\Pr(Y, X) \leq \min(\Pr(X), \Pr(Y))$.

Bayes' Rule

THE AUSTRALIAN NATIONAL UNIVERSITY

Conditional Probabilities, Part II

Obviously $\Pr(X, Y) = \Pr(Y, X)$. This allows us to write

$$\Pr(X|Y)\Pr(Y) = \Pr(X, Y) = \Pr(Y, X) = \Pr(Y|X)\Pr(X)$$

Solving for $\Pr(X|Y)$ yields Bayes' Rule

$$\Pr(X|Y) = \frac{\Pr(Y|X)\Pr(X)}{\Pr(Y)}$$

Interpretation

We can infer the probability of X, given Y, based on the individual probabilities of X, and Y and $\Pr(Y|X)$.

The usual application of Bayes' rule is that we use it to infer how likely a hypothesis is, given some experimental evidence.

Is the Pope an Alien?

See http://wol.ra.phy.cam.ac.uk/mackay/pope.html for details.

Examples

THE AUSTRALIAN NATIONAL UNIVERSITY

AIDS-Test

We want to find out likely it is that a patient *really* has AIDS (denoted by X) if the test is positive (denoted by Y).

Roughly 0.1% of all Australians are infected ($\Pr(X) = 0.001$). The probability that an AIDS test tells us the wrong result is in the order of 1% ($\Pr(Y|\mathcal{X}\backslash X) = 0.01$) and moreover we assume that it detects all infections ($\Pr(Y|X) = 1$). We have

$$\Pr(X|Y) = \frac{\Pr(Y|X)\Pr(X)}{\Pr(Y)} = \frac{\Pr(Y|X)\Pr(X)}{\Pr(Y|X)\Pr(X) + \Pr(Y|\mathcal{X}\backslash X)\Pr(\mathcal{X}\backslash X)}$$

Hence $\Pr(X|Y) = \frac{1 \cdot 0.001}{1 \cdot 0.001 + 0.01 \cdot 0.999} = 0.091$, i.e. the probability of AIDS is 9.1%!

Reliability of Eye-Witness

Assume that an eye-witness is 90% sure and that there were 20 people at the crime scene, what is the probability that the guy identified committed the crime?

$$\Pr(X|Y) = \frac{0.9 \cdot 0.05}{0.9 \cdot 0.05 + 0.1 \cdot 0.95} = 0.3213 = 32\% \text{ that's a worry } \ldots$$

Distribution, Measure, and Densities

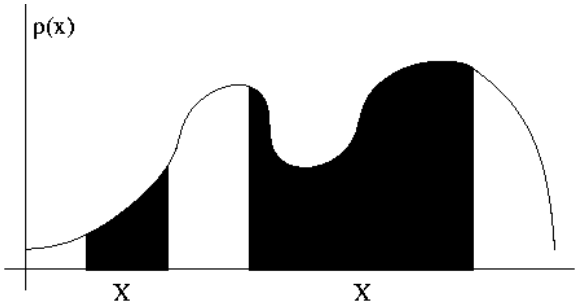
THE AUSTRALIAN NATIONAL UNIVERSITY

Computing $\Pr(X)$

Quite often $\mathcal{X}$ will not be a countable set, e.g., for $\mathcal{X} = \mathbb{R}$. So we need integrals.

$$\Pr(X) := \int_X d\Pr(x) = \int_X p(x)dx$$

Note that the last equality only holds if such a $p(x)$ exists. For the rest of this course we assume that such a p exists ...



Bayes' Rule for Densities

THE AUSTRALIAN NATIONAL UNIVERSITY

Multivariate Densities

As with conditional probabilities we can also have densities on product spaces $\mathcal{X} \times \mathcal{Y}$, given by $p(\mathbf{x}, \mathbf{y})$.

Conditional Densities

For independent variables the densities factorize and we have

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$$

For dependent variables (i.e. $\mathbf{x}$ tells us something about $\mathbf{y}$ and vice versa) we obtain

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$$

Bayes' Rule

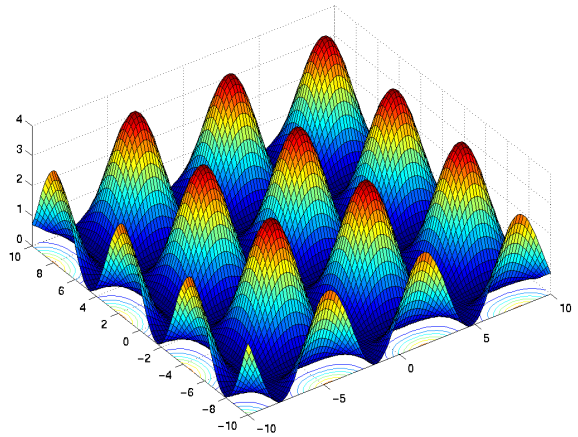
Solving for $p(\mathbf{y}|\mathbf{x})$ yields

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})}{p(\mathbf{x})}$$

Example: $p(\mathbf{x}) = 1 + \sin \mathbf{x}$

THE AUSTRALIAN NATIONAL UNIVERSITY

Factorizing Distribution



Random Variables

THE AUSTRALIAN NATIONAL UNIVERSITY

Definition

If we want to denote the fact that variables $\mathbf{x}$ and $\mathbf{y}$ are drawn at random from an underlying distribution, we call them *random variables*.

IID variables

This means Independent and Identically Distributed random variables. Most of our data has this property (e.g. each face we scan is independent of the other faces we saw). In this case the density factorizes into

$$p(\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}) = \prod_{i=1}^{m} p(\mathbf{x}_i)$$

Dependent Random Variables

For prediction purposes we want to estimate $\mathbf{y}$ from $\mathbf{x}$. In this case we *want* that $\mathbf{y}$ is *dependent* on $\mathbf{x}$. If $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$ we could not predict at all!

Marginalization and Conditioning

THE AUSTRALIAN NATIONAL UNIVERSITY

Marginalization

Given $p(\mathbf{x}, \mathbf{y})$ we want to obtain $p(\mathbf{x})$. For this purpose we integrate out $\mathbf{y}$ via

$$p(\mathbf{x}) = \int_{\mathcal{Y}} p(\mathbf{x}, \mathbf{y}) d\mathbf{y}$$

Conditioning

If we know $\mathbf{y}$, we can obtain $p(\mathbf{x}|\mathbf{y})$ via Bayes rule, i.e.

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{y})}.$$

A simler trick, however, is to note that the dependence of the RHS on $\mathbf{x}$ lies only in $p(\mathbf{x}, \mathbf{y})$ and therefore we obtain

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{\int_{\mathcal{X}} p(\mathbf{x}, \mathbf{y}) d\mathbf{x}}$$