DIGITAL TWIN AI and Machine Learning: Mathematical Foundations of Machine Learning

Prof. Andrew D. Bagdanov
andrew.bagdanov AT unifi.it

Dipartimento di Ingegneria dell'Informazione
Università degli Studi di Firenze

23 October 2020

Outline

Introduction

The mathematics of the 21st century

- Mastering contemporary machine learning requires a range of tools and disciplines.
- Our goal however is to just get up to speed on the basics.

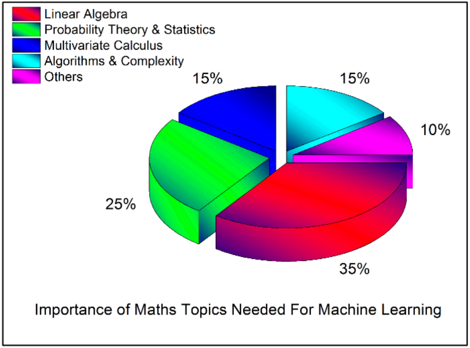


Image source: `https://towardsdatascience.com/the-mathematics-of-machine-learning-894f046c568`

Linear algebra

- Skyler Speakerman recently referred to Linear Algebra as the *mathematics for the 21st Century*.
- This might be slightly hyperbolic, but linear algebra is absolutely central to modern machine learning.
- Linear algebra allows us to deal with high dimensional data in a formal and precise way.
- It will allow us to model inputs to ML algorithms as points in high dimensional spaces.
- And subsequently to model functional transformations of these inputs into feature spaces.
- And finally, to model the subsequent transformations that lead to outputs (e.g. decisions or actions or estimates).

Linear algebra (continued)

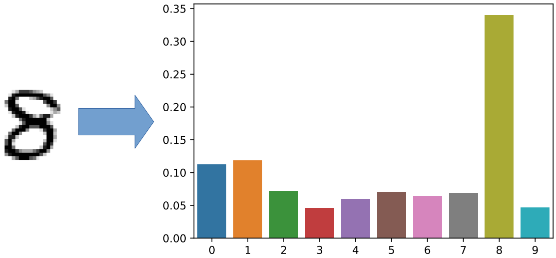
- What is an image? Is it a data structure, with width and height and depth, plus a corresponding array of raw data?
- We can go on... What is an audio recording? Or text document.
- Rather than define *ad hoc* data structures and algorithms, we want to treat them all the same.
- A 512×512 color image is a vector in a $512 \times 512 \times 3$-dimensional vector space.

Probability and statistics

- Perhaps somewhat surprisingly, probability and statistics are less important to modern machine learning.
- Sometimes we will want to give a probabilistic interpretation to a model or a model output.
- However, most deep learning models are defined as pure transformations of inputs into outputs.
- Often, these probabilistic interpretations are merely convenient fictions.
- Nonetheless, having a basic grasp of a few statistical concepts will be useful.
- As we will see, statistics and probability are much more useful as tools for analyzing results.

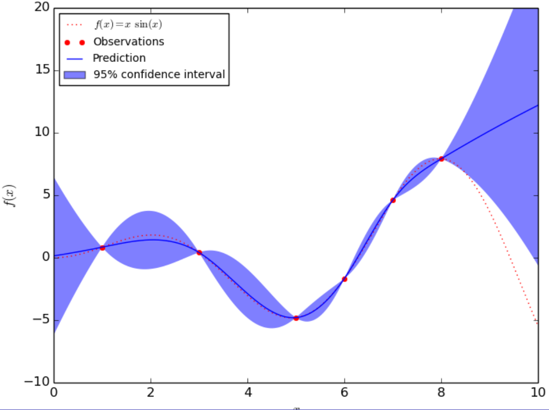
Probability and statistics (continued)

- For many problems we will want our models to output a probability distribution over possible outcomes.
- Take a simple classification problem: given an input image, estimate which digit is depicted.



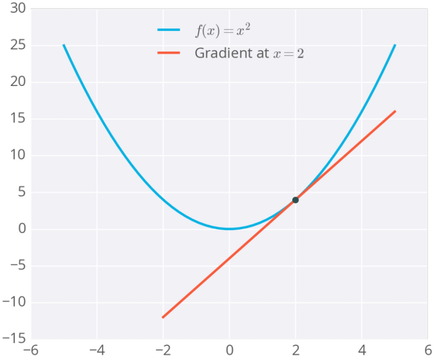
Probability and statistics

- For other problems we might want to qualify outputs of the model.
- This is the case in many regression problems where outputs at some points might be more certain than others.



Calculus and optimization

- Many (well, most) learning problems are formulated as optimization problems in (potentially very many) multiple variables.
- This means that to learn means to estimate these problems by minimizing some objective function.



Calculus and optimization (continued)

- For the most part the grisly details of numerical optimization will not concern us.
- We will rely on libraries and frameworks to take care of optimizing our objective functions.
- Nonetheless, it is useful to know what is happening when we fit a model to data.
- Typically, objective functions are highly non-convex (what does this mean?).
- Automatic differentiation and efficient algorithms like backpropagation (a clever implementation of the chain rule) come to the rescue here.

Numerical programming

- Tying everything together for this course will be practical, hand-on examples of many of the models we will study.
- These examples rely on tools like Numpy, scikit-learn, Tensorflow/Keras, and others.
- These tools were selected because they currently represent the best practices in academia and industry.
- While we will not concern ourselves with low-level details of the implementation, it is very useful to have a working knowledge of these numerical programming tools.

Numerical programming (example)

```
# Standard scientific Python imports
import matplotlib.pyplot as plt
from sklearn import datasets, svm, metrics
from sklearn.model_selection import train_test_split

# To apply a classifier on this data, we need to flatten the image, to
# turn the data in a (samples, feature) matrix:
digits = datasets.load_digits()
n_samples = len(digits.images)
data = digits.images.reshape((n_samples, -1))

# Create an SVM  classifier and split data into train/test.
classifier = svm.SVC(gamma=0.001)
X_train, X_test, y_train, y_test = train_test_split(
    data, digits.target, test_size=0.5, shuffle=False)

# We learn the digits on the first half of the digits
classifier.fit(X_train, y_train)
```

Preliminaries

Sets

- We are all familiar with the notion of a set, a collection of objects (the members of the set) without repetition.
- We can specify finite sets by enumerating their members: $E = \{0, 1\}$.
- We can also use the set former notation which defines sets as all elements satisfying a predicate: $E = \{x \mid P(x)\}$
- Set membership is indicated by $\in$: $x \in E$
- For example:

$$E = \{i \mid \text{i is an integer and there is an integer } j \text{ such that } i = 2j\}$$

- We will use quantifiers $\forall$ (for all/every) and $\exists$ (there exists) for conciseness:

$$E = \{i \mid i \in \mathbb{Z} \text{ and } \exists j \in \mathbb{Z} \text{ such that } i = 2j\}$$

- Question: what is the logical negation of $\forall$ and $\exists$?

Some useful sets and notation

Useful sets

- The universal set: $\mathbb{U}$ (needed sometimes, usually clear from context).
- The empty set: $\emptyset$ ($\forall x, x \notin \emptyset$).
- The integers: $\mathbb{Z}$ (the whole numbers).
- The natural numbers: $\mathbb{N}$ (non-negative integers).
- The real numbers: $\mathbb{R}$ (what we think of as numbers).
- Note: $\emptyset \subset \mathbb{N}^+ \subset \mathbb{N} \subset \mathbb{Z} \subset \mathbb{R}$.

Logical notation

- Quantifiers: $\forall$, $\exists$ (already seen).
- Operators: $p \wedge q$, $p \vee q$, $\neg p$ (p and q, p or q, not p).
- Implication: $\mathrm{p} \Rightarrow \mathrm{q}$ (if p then q).
- Equivalence: $p \Leftrightarrow q$ ($(p \Rightarrow q) \wedge (q \Rightarrow p)$, p iff q).

Operations and identities

Operations

- Complement: $A' = \{x \in \mathbb{U} \mid x \notin A\}$ $(= \overline{A})$
- Union: $A \cup B = \{x \mid x \in A \text{ or } x \in B\}$
- Intersection: $A \cap B = \{x \mid x \in A \text{ and } x \in B\}$
- Set difference: $A \setminus B = \{x \in A \mid x \notin B\}$
- Powerset: $\mathcal{P}(A) = \{B \mid B \subseteq A\}$ (a set of sets)
- Cartesian product: $A \times B = \{(a, b) \mid a \in A \text{ and } b \in B\}$

Identities

- Commutativity: $A \cup B = B \cup A, \;\; A \cap B = B \cap A$
- Associativity: $A \cup (B \cup C) = (A \cup B) \cup C$ (same for $\cap$)
- Distributivity: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ (reversed $\cup$ and $\cap$)
- De Morgen: $\overline{(A \cup B)} = A' \cap B'$

The Real Numbers

- A real number is a value of a continuous quantity that can represent a distance along a line.
- The real numbers include all the rational numbers, such as the integer -5 and the fraction 4/3, and all the irrational numbers, such as $\sqrt{2}$.
- Real numbers can be thought of as points on an infinitely long line called the number line or real line, where the points corresponding to integers are equally spaced.

Functions: basic definitions

- A function associates with each element of one set (the domain) a single element in another set (the codomain).
- If the function is f and the domain and codomain A and B, respectively, we write $f : A \to B$ to indicate that f is a function from A to B.
- For $f : A \to B$, we write $x \mapsto f(x)$ and say "f maps x to $f(x)$".
- We say $f : A \to B$ is injective (or is an injection) if whenever $f(x_1) = f(x_2)$, then $x_1 = x_2$.
- We say $f : A \to B$ is onto (or is surjective or a surjection) when $\forall b \in B, \exists a \in A$ s.t. $b = f(a)$.
- If $f : A \to B$ is injective and surjective, we say that it is one-to-one or that it is bijective.

Cartesian products

- When we write $\mathbb{R} \times \mathbb{R}$ we are referring to the set of pairs of real numbers:

$$\mathbb{R} \times \mathbb{R} = \{\, (x, y) \;\mid\; x \in \mathbb{R} \text{ and } y \in \mathbb{R} \,\}$$

- Which we can naturally generalize to arbitrary dimensions:

$$\mathbb{R}^n = \{\, (x_1, x_2, \ldots, x_n) \;\mid\; x_i \in \mathbb{R} \text{ for } 1 \leq i \leq n \,\}$$

- This will let us compactly define functions of multiple arguments which return multiple arguments:

$$\begin{aligned} f : \mathbb{R}^2 &\rightarrow \mathbb{R} \\ f(\mathbf{x}) &= \mathbf{x}^T \mathbf{x} \end{aligned}$$

Linear Algebra

Vectors and vector spaces

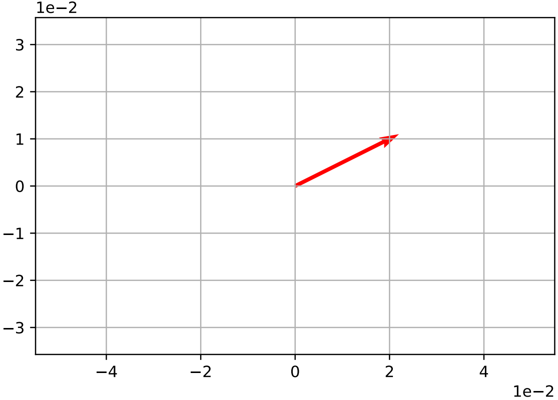
- Vectors and vector spaces are fundamental to linear algebra.
- Vectors describe lines, planes, and hyperplanes in space.
- They allow us to perform calculations that explore relationships in multi-dimensional spaces.
- At its simplest, a vector is a mathematical object that has both magnitude and direction.
- We write vectors using a variety of notations, but we will usually write them like this:

$$\mathbf{v} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

- The boldface symbol lets us know it is a vector.

Vectors and vector spaces (continued)

- What does it mean to have direction and magnitude?
- Well, it helps to look at a visualization (in at most three dimensions):



Vectors and vector spaces (continued)

More formally, we say that $\mathbf{v}$ is a vector in n dimensions (or rather, $\mathbf{v}$ is a vector in the vector space $\mathbb{R}^n$) if:

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

for $v_i \in \mathbb{R}$. Note that we use regular symbols (i.e. not boldfaced) to refer to the individual elements of $\mathbf{v}$.

Operations on vectors

Definition (Fundamental vector operations)

- Vector addition: if $\mathbf{u}$ and $\mathbf{v}$ are vectors in $\mathbb{R}^n$, then so is $\mathbf{w} = \mathbf{u} + \mathbf{v}$ (where we define $w_i = u_i + v_i$).
- Scalar multiplication: if $\mathbf{v}$ is a vector in $\mathbb{R}^n$, then so is $\mathbf{w} = c\mathbf{v}$ for any $c \in \mathbb{R}$ (we define $w_i = cv_i$).
- Scalar (dot) product: if $\mathbf{u}$ and $\mathbf{v}$ are vectors in $\mathbb{R}^n$, we define the scalar or dot product as:

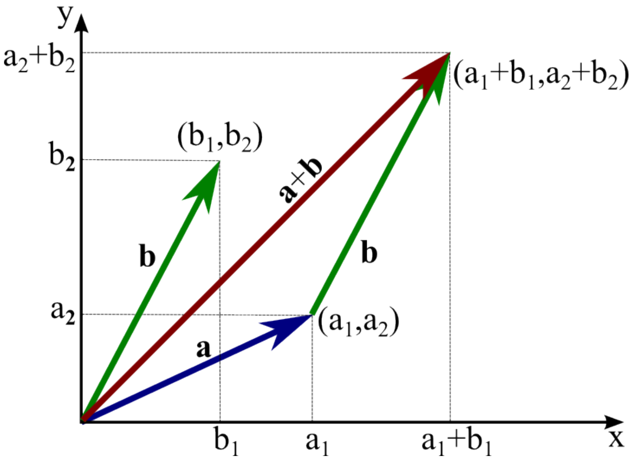
$$\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^{n} u_i v_i$$

- Vector norm (or magnitude, or length): if $\mathbf{v}$ is a vector in $\mathbb{R}^n$, then we define the norm or length of $\mathbf{v}$ as:

$$||\mathbf{u}|| = \sqrt{\mathbf{u} \cdot \mathbf{u}}$$

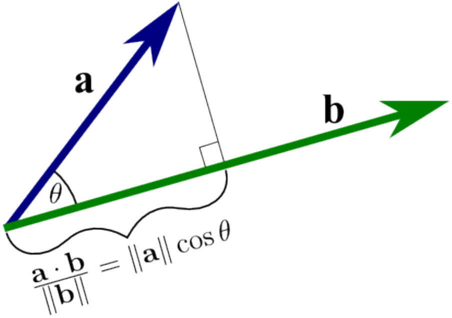
Visualizing vectors (in 2D)

- Vector addition is easy to interpret in 2D:



Visualizing the dot product

- The scalar or dot product is related to the directions and magnitudes of the two vectors:



- In fact, it is easy to recover the cosine between any two vectors.
- Note that these properties generalize to any number of dimensions.
- Question: how can we test it two vectors are perpendicular (orthogonal)?

Matrices: basics

- A matrix arranges numbers into rows and columns, like this:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

- Note that matrices are generally named as a capital, boldface letter. We refer to the elements of the matrix using the lower case equivalent with a subscript row and column indicator:

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \end{bmatrix}$$

- Here we say that $\mathbf{A}$ is a matrix of size 2×3.
- Equivalently: $\mathbf{A} \in \mathbb{R}^{2\times 3}$.

Matrices: arithmetic operations

- Matrices support common arithmetic operations:
- To add two matrices of the same size together, just add the corresponding elements in each matrix:
$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} + \begin{bmatrix} 6 & 5 & 4 \\ 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 7 & 7 & 7 \\ 7 & 7 & 7 \end{bmatrix}$$
- Each matrix has two rows of three columns (so we describe them as 2×3 matrices).
- Adding matrices $\mathbf{A} + \mathbf{B}$ results in a new matrix $\mathbf{C}$ where $c_{i,j} = a_{i,j} + b_{i,j}$.
- This *elementwise* definition generalizes to subtraction, multiplication and division.

Matrices: arithmetic operations (continued)

- In the previous examples, we were able to add and subtract the matrices, because the operands (the matrices we are operating on) are conformable for the specific operation (in this case, addition or subtraction).
- To be conformable for addition and subtraction, the operands must have the same number of rows and columns
- There are different conformability requirements for other operations, such as multiplication.

Matrices: unary arithmetic operations

- The negation of a matrix is just a matrix with the sign of each element reversed:

$$C = \begin{bmatrix} -5 & -3 & -1 \\ 1 & 3 & 5 \end{bmatrix}$$

$$-C = \begin{bmatrix} 5 & 3 & 1 \\ -1 & -3 & -5 \end{bmatrix}$$

- The transpose of a matrix switches the orientation of its rows and columns.
- You indicate this with a superscript T, like this:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

Matrices: matrix multiplication

- Multiplying matrices is a little more complex than the elementwise arithmetic we have seen so far.
- There are two cases to consider, scalar multiplication (multiplying a matrix by a single number)

$$2 \times \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 6 \\ 8 & 10 & 12 \end{bmatrix}$$

- And dot product matrix multiplication:

$$\mathbf{AB} = \mathbf{C}, \text{ where } c_{i,j} = \sum_{k=1}^{n} a_{i,k} b_{k,j}$$

- What can we infer about the conformable sizes of **A** and **B**? What is the size of **C**.

Matrices: multiplication is just dot products

- To multiply two matrices, we are really calculating the dot product of rows and columns.
- We perform this operation by applying the RC rule - always multiplying (dotting) Rows by Columns.
- For this to work, the number of columns in the first matrix must be the same as the number of rows in the second matrix so that the matrices are conformable.
- An example:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \cdot \begin{bmatrix} 9 & 8 \\ 7 & 6 \\ 5 & 4 \end{bmatrix} = \begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}$$

Matrices: inverses

- The identity matrix $\mathbf{I}$ is a square matrix with all ones on the diagonal, and zeros everywhere else.
- So, $\mathbf{IA} = \mathbf{BI}$, and $\mathbf{Iv} = \mathbf{v}$.
- The inverse of a square matrix $mathbf A$ is denoted $\mathbf{A}^{-1}$.
- $\mathbf{A}^{-1}$ is the unique (if it exists) matrix such that:

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

Matrices: solving systems of equations

- We can now use this to our advantage:

$$\begin{bmatrix} 67.9 & 1.0 \\ 61.9 & 1.0 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} 170.85 \\ 122.50 \end{bmatrix}$$

- Multiplying both sides by the inverse:

$$\begin{bmatrix} 67.9 & 1.0 \\ 61.9 & 1.0 \end{bmatrix}^{-1} \begin{bmatrix} 67.9 & 1.0 \\ 61.9 & 1.0 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} 67.9 & 1.0 \\ 61.9 & 1.0 \end{bmatrix}^{-1} \begin{bmatrix} 170.85 \\ 122.50 \end{bmatrix}$$

- And we have:

$$\mathbf{I} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} 67.9 & 1.0 \\ 61.9 & 1.0 \end{bmatrix}^{-1} \begin{bmatrix} 170.85 \\ 122.50 \end{bmatrix}$$

Matrices: linear versus affine

- Matrix multiplication computes linear transformations of vector spaces.
- We are also interested in affine transformations that don't necessarily preserve the origin:
- An affine transformation is a linear transformation followed by a translation:

$$f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$$

- Note: an affine transformation in n dimensions can be modeled by a linear transformation in $n + 1$ dimensions.

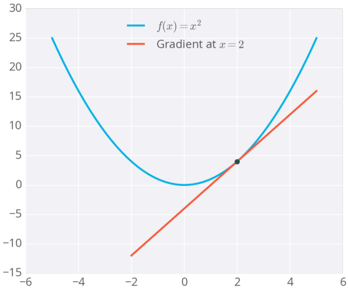
A general structure for dense data

- There is nothing magic about one dimension (vectors) or two dimensions (matrices).
- In fact, the tools we use are completely generic in that we can define dense, homogeneous arrays of numeric data of any dimensionality.
- The generic term for this is a tensor, and all of the math generalizes to arbitrary dimensions.
- Example: a color image is naturally modeled as a tensor in three dimensions (two spatial, one chromatic).
- Example: a batch of b color images of size 32×32 is easily modeled by simply adding a new dimension: $\mathbf{B} \in \mathbb{R}^{b \times 32 \times 32 \times 3}$.

Calculus and Optimization

Return to our illustrative example

- Let's say we want to find the minimal value of the function $f(x) = x^2$.
- Here's a recipe:
  1. Start with an initial guess x_0.
  2. Take a small step in the direction of steepest descent; call this x_{i+1}.
  3. If $|f(x_{i+1}) - f(x_i)| < \varepsilon$, stop.
  4. Otherwise: repeat from 2.



Gradient descent

- Maybe the only thing imprecise about this recipe is the definition of small step in the direction of steepest descent.
- Well, in one variable we know how to do this:

$$x_{i+1} = x_i - \eta \frac{d}{dx} f(x_i)$$

- So the derivative gives us the direction, and the parameter η defines what "small" means.
- This recipe also works in more dimensions:

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_i)$$

- Let's dissect this. . .

Fitting models with gradient descent

- Many of the models we will see have a form like:

$$f(\mathbf{x}; \boldsymbol{\theta}) : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

- That is: function f is parameterized by parameters $\boldsymbol{\theta}$.
- Goal: find a $\boldsymbol{\theta}^*$ that optimize some fitness criterion $\mathcal{L}$ on data $\mathbf{D}$:

$$\boldsymbol{\theta}^* = \arg\min \mathcal{L}(\mathbf{D}, \boldsymbol{\theta})$$

- Example (least squares):

$$\begin{aligned} D &= \{ (x_i, y_i) \mid 1 \leq i \leq n \} \\ \boldsymbol{\theta} &= \begin{bmatrix} m \\ b \end{bmatrix} \\ \mathcal{L}(D, \boldsymbol{\theta}) &= \sum_i ||(mx_i + b) - y_i||_2 \end{aligned}$$

Statistics

Discrete probability distributions

To specify a discrete random variable, we need a sample space and a probability mass function:

- Sample space Ω: Possible states x of the random variable X (outcomes of the experiment, output of the system, measurement).
- Discrete random variables have a finite number of states.
- Events: Possible combinations of states (subsets of Ω)
- Probability mass function $P(X = x)$: A function which tells us how likely each possible outcome is:

$$P(X = x) = P_X(x) = P(x)$$

$$P(x) \geq 0 \text{ for each } x$$

$$\sum_{x \in \Omega} P(x) = 1$$

$$P(A) = P(x \in A) = \sum_{x \in A} P(X = x)$$

Discrete probability distributions (continued)

- Conditional probability: Recalculated probability of event A after someone tells you that event B happened:

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

- Example: rolling dice [on board]
- Bayes Rule:

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

Discrete probability distributions (continued)

Expectation and variance characterize the mean value of a random variable and its dispersion:

- Expectation: $E(X) = \sum_x P(X = x)x$
- Expectation of a function: $E(f(X)) = \sum_x P(X = x)f(x)$
- Moments: the expectation of a power of X: $M_k = E(X^k)$
- Variance: Average (squared) fluctuation from the mean:

$$\begin{aligned} \mathrm{Var}(X) &= E((X - E(X))^2) \\ &= E(X^2) - E(X)^2 \\ &= M_2 - M_1^2 \end{aligned}$$

- Standard deviation: square root of variance.
- Aside: Difference between expectation/variance of random variable and empirical average/variance.

Multivariate probability distributions

Bivariate distributions characterize systems with two observables:

- Joint distribution: $P(X = x, Y = y)$, a list of all probabilities of all possible pairs of observations.
- Marginal distribution: $P(X = x) = \sum_y P(X = x, Y = y)$
- Conditional distribution: $P(X = x|Y = y) = \frac{P(X=x,Y=y)}{P(y=y)}$
- $X|Y$ has distribution $P(X|Y)$, where $P(X|Y)$ specifies a 'lookup-table' of all possible $P(X = x|Y = y)$.

Conditioning and marginalization come up in Bayesian inference ALL the time: *Condition on what you observe, Marginalize out the uncertainty.*

Expectation and covariance of multivariate distributions

- Conditional distributions are just distributions which have a (conditional) mean or variance.
- Note: $E(X|Y) = f(Y)$ – *If I tell you what Y is, what is the average value of X?*
- Covariance is the expected value of the product of fluctuations:

$$\mathrm{Cov}(X, Y) = E\left((X - E(X))(Y - E(Y))\right) \tag{1}$$
$$= E(XY) - E(X)E(Y) \tag{2}$$
$$\mathrm{Var}(X) = \mathrm{Cov}(X, X) \tag{3}$$

Independence of random variables

- Intuitively, two events are independent if knowing that the first took places tells us nothing about the probability of the second: $P(A|B) = P(A)$ $(P(A)P(B) = P(A \cap B))$.
- If X and Y are independent, we write $X \perp Y$: knowing the value of X does not tell us anything about Y.
- If X and Y are independent, $\text{Cov}(X, Y) = 0$.

Multivariate distributions: the same, but different

- Multivariate distributions are the same as bivariate distributions – just with more dimensions.
- $\mathbf{X}, \mathbf{x}$ are vector valued.
- Mean: $E(\mathbf{X}) = \sum_{\mathrm{x}} \mathbf{x} P(\mathbf{x})$
- Covariance matrix:

$$\begin{aligned} \mathrm{Cov}(X_i, X_j) &= E(X_i X_j) - E(X_i)E(X_j) \\ \mathrm{Cov}(\mathbf{X}) &= E(\mathbf{X}\mathbf{X}^\top) - E(\mathbf{X})E(\mathbf{X})^\top \end{aligned}$$

- Conditional and marginal distributions: Can define and calculate any (multi or single-dimensional) marginals or conditional distributions we need: $P(X_1)$, $P(X_1, X_2)$, $P(X_1, X_2, X_3 | X_4)$, etc..

Continuous random variables

- A random variable X is continuous if its sample space X is uncountable.
- In this case, $P(X = x) = 0$ for each x (measure zero support).
- If $p_X(x)$ is a probability density function for X, then:

$$P(a < X < b) = \int_a^b p(x)dx$$

- The cumulative distribution function is $F_X(x) = P(X < x)$. We have that $p_X(x) = F'(x)$, and $F(x) = \int_{-\infty}^x p(s)ds$.
- More generally: If A is an event, then

$$P(A \subseteq \Omega) = P(X \in A) = \int_{x \in A} p(x)dx$$

$$P(\Omega) = P(X \in \Omega) = \int_{x \in \Omega} p(x)dx = 1$$

Probability, mass and density

- People (including me) will often say probability when they mean probability density.
- Probability density functions do not satisfy the definitions of probability (e.g. they can bigger than 1).
- However, people will often be sloppy and write things like $P(X = x)$ and say 'the probability of X' when they really mean 'the probability density of X evaluated at x'.
- This might be bad practice, but it is usually clear from the context whether a random variable is discrete or continuous.
- In addition, it is good preparation for reading papers — many machine learning papers are very sloppy about usage of these terms.

Mean, variance, and conditioning of continuous RVs

- Mostly the same as the discrete case, just with sums replaced by integrals.
- Mean: $E(X) = \int_x xp(x)dx$
- Variance: $\mathrm{Var}(X) = E(X^2) - E(X)^2$
- Conditioning: If X has pdf $p(x)$, then $X|(X \in A)$ has pdf:

$$p_{X|A}(x) = \frac{p(x)}{P(A)} = \frac{p(x)}{\int_{x \in A} p(x)dx}$$

Conditioning and independence of continuous random variables

- $p_{X,Y}(x, y) = p(x, y)$, joint probability density function of X and Y.
- $\int_x \int_y p(x, y) dx dy = 1$
- Marginal distribution: $p(x) = \int_{-\infty}^{\infty} p(x, y) dy$
- Conditional distribution $p(x|y) = \frac{p(x,y)}{p(y)}$
- Note: $P(Y = y) = 0$! Formally, conditional probability in the continuous case can be derived using infinitesimal events.
- Independence: X and Y are independent if $p(x, y) = p(x)p(y)$

The univariate Gaussian (normal) distribution

- The Univariate Gaussian:

$$t \sim \mathcal{N}(\mu, \sigma^2)$$
$$p(t|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right)$$

- The Gaussian has mean μ and variance σ^2 and precision $\beta = 1/\sigma^2$
- What are the mode and the median of the Gaussian?

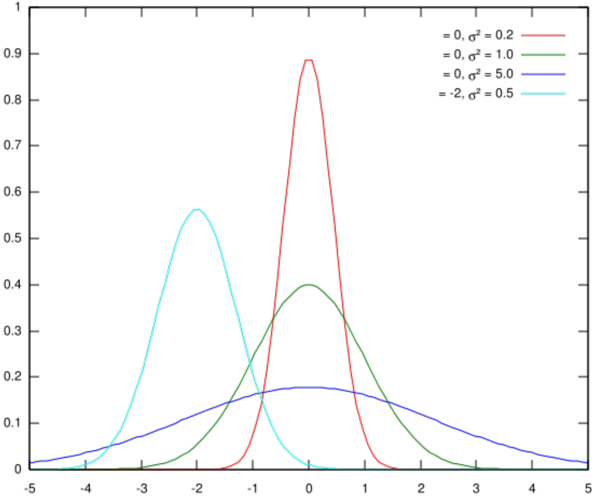
Products of Gaussians

- An aside: products of Gaussian pdfs are (unnormalized) Gaussians pdfs.
- Suppose $p_1(x) = \mathcal{N}(x, \mu_1, 1/\beta_1)$ and $p_2(x) = \mathcal{N}(x, \mu_2, 1/\beta_2)$, then:

$$
\begin{aligned}
p_1(x)p_2(x) &\propto \mathcal{N}(x, \mu, 1/\beta) \\
\beta &= \beta_1 + \beta_2 \\
\mu &= \frac{1}{\beta}(\beta_1\mu_1 + \beta_2\mu_2)
\end{aligned}
$$

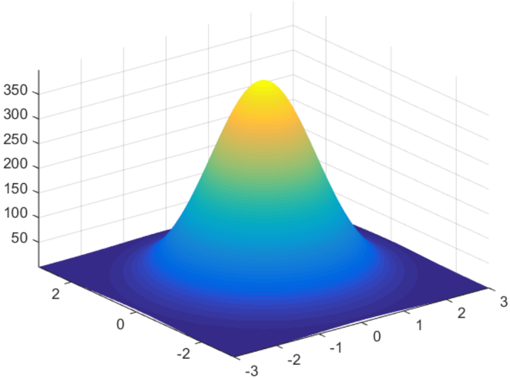
Gaussian distributions

- As they say, a picture is worth a thousand words:



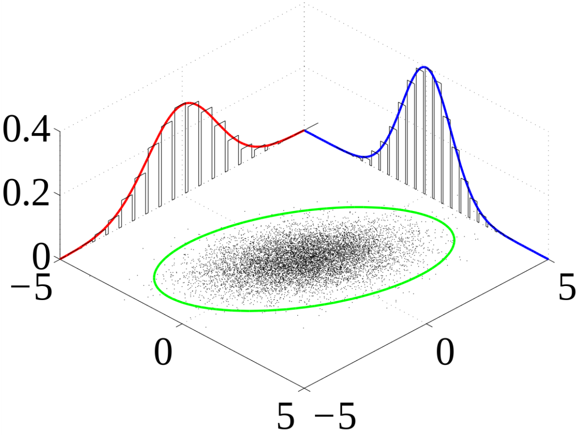
The multivariate Gaussian

$$f(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}))$$



The multivariate Gaussian (marginals)

- An important property of the multivariate Gaussian is that its marginals are also Gaussian:



Reflections

Mathematical tools of the trade

- Most of the details and abstract properties of gradients, matrices, tensors, et al., are not terribly important.
- Mostly, these tools are useful as a working vocabulary.
- They will allow us to formulate learning problems using this common vocabulary – which is already useful just as a communication tool.
- More importantly: formulating problems in this language allows us to communicate with the tools we will use to fit models.

Foundations and Numpy Lab

- OK, now we can go to this URL for today's lab:

http://bit.ly/DTwin-ML2