

# CS 294-34: Practical Machine Learning

## Tutorial

Ariel Kleiner

Content inspired by Fall 2006  
tutorial lecture by Alexandre Bouchard-Cote and Alex Simma

January 22, 2008

# Machine Learning Draws Heavily On...

- Probability and Statistics
- Optimization
- Algorithms and Data Structures

# Probability: Foundations

A probability space  $(\Omega, \mathcal{F}, P)$  consists of

- a set  $\Omega$  of "possible outcomes"
- a set<sup>1</sup>  $\mathcal{F}$  of events, which are subsets of  $\Omega$
- a probability measure  $P : \mathcal{F} \rightarrow [0, 1]$  which assigns probabilities to events in  $\mathcal{F}$

## Example: Coin-Flipping

Consider a coin that lands on heads with probability  $p \in [0, 1]$ . In this case,

$$\Omega = \{H, T\}$$

$$\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$$

$$P(\emptyset) = 0, P(\{H\}) = p, P(\{T\}) = 1 - p, P(\{H, T\}) = 1$$

---

<sup>1</sup>Actually,  $\mathcal{F}$  is a  $\sigma$ -field. See Durrett's *Probability: Theory and Examples* for thorough coverage of the measure-theoretic basis for probability theory.

# Probability: Random Variables

- A random variable is an assignment of (often numeric) values to outcomes in  $\Omega$ .
- For a set  $A$  in range of a random variable  $X$ , the probability that  $X$  falls in  $A$  is written as  $P(X \in A)$ . We have thus induced a probability distribution for the random variable.

## Example Continued: Coin-Flipping

Suppose that we bet \$1 on each coin flip. Let  $X : \{H, T\} \rightarrow \{-1, 1\}$  be a random variable denoting our winnings:  $X = 1$  if the coin lands on heads, and  $X = -1$  if the coin lands on tails. Furthermore,

$$P(X \in \{1\}) = p, P(X \in \{-1\}) = 1 - p.$$

# Probability: Common Discrete Distributions

Various distributions arise commonly and are useful in both theory and application. The following are among the most common discrete distributions for a random variable  $X$ :

- Bernoulli( $p$ ):  $p \in [0, 1]$ ;  $X \in \{0, 1\}$

$$P(X = 1) = p, P(X = 0) = 1 - p$$

- Binomial( $p, n$ ):  $p \in [0, 1]$ ,  $n \in \mathbb{N}$ ;  $X \in \{0, \dots, n\}$

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

- The multinomial distribution generalizes the Bernoulli and the Binomial beyond binary outcomes for individual experiments.
- Poisson( $\lambda$ ):  $\lambda \in (0, \infty)$ ;  $X \in \mathbb{N}$

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

## Probability: More on Random Variables

- Notation:  $X \sim P$  means "X has the distribution given by  $P$ "
- The cumulative distribution function (cdf) of a random variable  $X \in \mathbb{R}^m$  is defined for  $x \in \mathbb{R}^m$  as  $F(x) = P(X \leq x)$ .
- We say that  $X$  has a density function  $p$  if we can write  $P(X \leq x) = \int_{-\infty}^x p(y)dy$  (densities exist under certain regularity conditions).
- In practice, the continuous random variables with which we will work will have densities.
- For convenience, in the remainder of this lecture we will assume that all random variables take values in some countable numeric set,  $\mathbb{R}$ , or a real vector space.

# Probability: Common Continuous Distributions

The following are among the most common continuous distributions for a random variable  $X$  (all have densities  $p$ ):

- Uniform( $a, b$ ):  $a, b \in \mathbb{R}$ ,  $a < b$ ;  $X \in [a, b]$

$$p(x) = \frac{1}{b - a}$$

- Normal( $\mu, \sigma^2$ ):  $\mu \in \mathbb{R}$ ,  $\sigma \in \mathbb{R}_{++}$ ;  $X \in \mathbb{R}$

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

- The normal distribution can be easily generalized to the multivariate case, in which  $X \in \mathbb{R}^m$ . In this context,  $\mu$  becomes a real vector and  $\sigma$  is replaced by a covariance matrix.
- Other continuous distributions that also frequently arise are the Beta, Gamma, and Dirichlet.

# Probability: Distributions

## Other Distribution Types

### Exponential Family

- encompasses distributions of the form

$$P(X = x) = h(x) \exp(\eta(\theta)T(x) - A(\theta))$$

- includes many commonly encountered distributions
- well-studied and has various nice analytical properties while being fairly general

### Graphical Models

Graphical models provide a flexible framework for building complex models involving many random variables while allowing us to leverage conditional independence relationships among them to control computational tractability.

# Probability: Expectation

- *Intuition*: the expectation of a random variable is its "average" value under its distribution.
- Formally, the expectation of a random variable  $X$ , denoted  $E[X]$ , is its Lebesgue integral with respect to its distribution.
- If  $X$  takes values in some countable numeric set  $\mathcal{X}$ , then

$$E[X] = \sum_{x \in \mathcal{X}} xP(X = x)$$

.

- If  $X \in \mathbb{R}^m$  has a density  $p$ , then

$$E[X] = \int_{\mathbb{R}^m} xp(x)dx$$

.

- These ideas generalize naturally to expectations of functions of random variables.

# Probability: More on Expectation

- Expectation is linear:  $E[aX + b] = aE[X] + b$ . Also, if  $Y$  is also a random variable, then  $E[X + Y] = E[X] + E[Y]$ .
- Expectation is monotone: if  $X \geq Y$ , then  $E[X] \geq E[Y]$ .
- Expectations also obey various inequalities, including Jensen's, Cauchy-Schwarz, and Chebyshev's.

## Variance

The variance of a random variable  $X$  is defined as

$$\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

and obeys the following for  $a, b \in \mathbb{R}$ :

$$\text{Var}(aX + b) = a^2 \text{Var}(X).$$

# Probability: Independence

- *Intuition*: two random variables are independent if knowing the value of one yields no knowledge about the value of the other.
- Formally, two random variables  $X$  and  $Y$  are independent iff  $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$  for all (measurable) subsets  $A$  and  $B$  in the ranges of  $X$  and  $Y$ .
- If  $X, Y$  have densities  $p_X(x), p_Y(y)$ , then they are independent if  $p_{X,Y}(x, y) = p_X(x)p_Y(y)$ .

# Probability: Conditioning

- *Intuition*: conditioning allows us to capture the probabilistic relationships between different random variables.
- For events  $A$  and  $B$ ,  $P(A|B)$  is the probability that  $A$  will occur given that we know that event  $B$  has occurred. If  $P(B) > 0$ , then

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

- If  $X, Y$  have joint density  $p(x, y)$ , the conditional density of  $X$  given  $Y$  is as follows:

$$p(y|x) = \frac{p(x, y)}{p(x)}, \text{ for } p(x) > 0$$

where  $p(x) = \int p(x, y) dy$ .

- If  $X$  and  $Y$  are independent, then  $P(Y = y|X = x) = P(Y = y)$  and  $P(X = x|Y = y) = P(X = x)$ .

## Probability: More on Conditional Probability

- For any events  $A$  and  $B$  (e.g., we might have  $A = \{Y \leq 5\}$ ),

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

- *Bayes' Theorem:*

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

Equivalently, if  $P(B) > 0$ ,

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

- Bayes' Theorem provides a means of inverting the "order" of conditioning.

# Probability: Law of Large Numbers

## Strong Law of Large Numbers

Let  $X_1, X_2, X_3, \dots$  be independent identically distributed (i.i.d.) random variables with  $E|X_j| < \infty$ . Then

$$\frac{1}{n} \sum_{i=1}^n X_i \rightarrow E[X_1]$$

with probability 1 as  $n \rightarrow \infty$ .

## Application: Monte Carlo Methods

How can we compute an (approximation of) an expectation  $E[f(X)]$  with respect to some distribution  $P$  of  $X$ ? (assume that we can draw independent samples from  $P$ ).

*A Solution:* Draw a large number of samples  $x_1, \dots, x_n$  from  $P$ .

Compute  $E[f(X)] \approx \frac{f(x_1) + \dots + f(x_n)}{n}$ .

# Probability: Central Limit Theorem

- The Central Limit Theorem provides insight into the distribution of a normalized sum of independent random variables. In contrast, the law of large numbers only provides a single limiting value.
- *Intuition:* The sum of a large number of small, independent, random terms is asymptotically normally distributed.
- This theorem is heavily used in statistics.

## Central Limit Theorem

Let  $X_1, X_2, X_3, \dots$  be i.i.d. random variables with  $E[X_i] = \mu$ ,  $\text{Var}(X_i) = \sigma^2 \in (0, \infty)$ . Then, as  $n \rightarrow \infty$ ,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mu}{\sigma} \xrightarrow{d} N(0, 1)$$

# Statistics: Frequentist Basics

- Given data (i.e., realizations of random variables)  $x_1, x_2, \dots, x_n$  which is generally assumed to be i.i.d.
- Based on this data, we would like to estimate some (unknown) value  $\theta$  associated with the distribution from which the data was generated.
- In general, our estimate will be a function  $\hat{\theta}(x_1, \dots, x_n)$  of the data (i.e., a statistic).

## Examples

- Given the results of  $n$  independent flips of a coin, determine the probability  $p$  with which it lands on heads.
- Simply determine whether or not the coin is fair.
- Find a function that distinguishes digital images of fives from those of other handwritten digits.

# Statistics: Parameter Estimation

- In practice, we often seek to select from some class of distributions a single distribution corresponding to our data.
- If our model class is parametrized by some (possibly uncountable) set of values, then this problem is that of parameter estimation.
- That is, from a set of distributions  $\{p_\theta(x) : \theta \in \Theta\}$ , we will select that corresponding to our estimate  $\hat{\theta}(x_1, \dots, x_n)$  of the parameter.
- How can we obtain estimators in general?
- *One answer:* maximize the likelihood  
 $l(\theta; x_1, \dots, x_n) = p_\theta(x_1, \dots, x_n) = \prod_{i=1}^n p_\theta(x_i)$  (or, equivalently, log likelihood) of the data.

## Maximum Likelihood Estimation

$$\hat{\theta}(x_1, \dots, x_n) = \operatorname{argmax}_{\theta \in \Theta} \prod_{i=1}^n p_\theta(x_i) = \operatorname{argmax}_{\theta \in \Theta} \sum_{i=1}^n \ln p_\theta(x_i)$$

# Statistics: Maximum Likelihood Estimation

## Example: Normal Mean

- Suppose that our data is real-valued and known to be drawn i.i.d. from a normal distribution with variance 1 but unknown mean.
- *Goal*: estimate the mean  $\theta$  of the distribution.
- Recall that a univariate  $N(\theta, 1)$  distribution has density  $p_\theta(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(x - \theta)^2)$ .
- Given data  $x_1, \dots, x_n$ , we can obtain the maximum likelihood estimate by maximizing the log likelihood w.r.t.  $\theta$ :

$$\frac{d}{d\theta} \sum_{i=1}^n \ln p_\theta(x_i) = \sum_{i=1}^n \frac{d}{d\theta} \left[ -\frac{1}{2}(x_i - \theta)^2 \right] = \sum_{i=1}^n (x_i - \theta) = 0$$

$$\Rightarrow \hat{\theta}(x_1, \dots, x_n) = \operatorname{argmax}_{\theta \in \Theta} \sum_{i=1}^n \ln p_\theta(x_i) = \frac{1}{n} \sum_{i=1}^n x_i$$

# Statistics: Criteria for Estimator Evaluation

- *Bias*:  $B(\theta) = E_{\theta}[\hat{\theta}(X_1, \dots, X_n)] - \theta$
- *Variance*:  $\text{Var}_{\theta}(\hat{\theta}(X_1, \dots, X_n)) = E_{\theta}[(\hat{\theta} - E_{\theta}[\hat{\theta}])^2]$
- *Loss/Risk*
  - A loss function  $L(\theta, \hat{\theta}(X_1, \dots, X_n))$  assigns a penalty to an estimate  $\hat{\theta}$  when the true value of interest is  $\theta$ .
  - The risk is the expectation of the loss function:  
 $R(\theta) = E_{\theta}[L(\theta, \hat{\theta}(X_1, \dots, X_n))]$ .
  - *Example*: squared loss is given by  $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ .

## Bias-Variance Decomposition

Under squared loss,

$$E_{\theta}[L(\theta, \hat{\theta})] = E_{\theta}[(\theta - \hat{\theta})^2] = [B(\theta)]^2 + \text{Var}_{\theta}(\hat{\theta})$$

- *Consistency*: Does  $\hat{\theta}(X_1, \dots, X_n) \xrightarrow{P} \theta$  as  $n \rightarrow \infty$ ?

# Statistics: Criteria for Estimator Evaluation

## Example: Evaluation of Maximum Likelihood Normal Mean Estimator

Recall that, in this example,  $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} N(\theta, 1)$  and the maximum likelihood estimator for  $\theta$  is

$$\hat{\theta}(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i$$

Therefore, we have the following:

- *Bias:*

$$\begin{aligned} B(\theta) &= E_{\theta}[\hat{\theta}(X_1, \dots, X_n)] - \theta = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] - \theta \\ &= \frac{1}{n} \sum_{i=1}^n E[X_i] - \theta = \frac{1}{n} \sum_{i=1}^n \theta - \theta = 0 \end{aligned}$$

- *Variance:*  $\text{Var}(\hat{\theta}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n}$
- *Consistency:*  $\frac{1}{n} \sum_{i=1}^n X_i \rightarrow E[X_1] = \theta$  with probability 1 as  $n \rightarrow \infty$ , by the strong law of large numbers.

# Statistics: Bayesian Basics

- The Bayesian approach treats statistical problems by maintaining probability distributions over possible parameter values.
- That is, we treat the parameters themselves as random variables having distributions:
  - 1 We have some beliefs about our parameter values  $\theta$  before we see any data. These beliefs are encoded in the *prior distribution*  $P(\theta)$ .
  - 2 Treating the parameters  $\theta$  as random variables, we can write the likelihood of the data  $X$  as a conditional probability:  $P(X|\theta)$ .
  - 3 We would like to update our beliefs about  $\theta$  based on the data by obtaining  $P(\theta|X)$ , the *posterior distribution*.  
*Solution:* by Bayes' theorem,

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

where

$$P(X) = \int P(X|\theta)P(\theta)d\theta$$

# Statistics: More on the Bayesian Approach

- Within the Bayesian framework, estimation and prediction simply reduce to probabilistic inference. This inference can, however, be analytically and computationally challenging.
- It is possible to obtain point estimates from the posterior in various ways, such as by taking the posterior mean

$$E_{\theta|X}[\theta] = \int \theta P(\theta|X) d\theta$$

or the mode of the posterior:

$$\operatorname{argmax}_{\theta} P(\theta|X)$$

- Alternatively, we can directly compute the predictive distribution of a new data point  $X_{\text{new}}$ , having already seen data  $X$ :

$$P(X_{\text{new}}|X) = \int P(X_{\text{new}}|\theta)P(\theta|X)d\theta$$

# Statistics: Bayesian Approach for the Normal Mean

Suppose that  $X|\theta \sim N(\theta, 1)$  and we place a prior  $N(0, 1)$  over  $\theta$  (i.e.,  $\theta \sim N(0, 1)$ ):

$$P(X = x|\theta) = \frac{1}{2\pi} \exp\left(-\frac{(x - \theta)^2}{2}\right) \quad P(\theta) = \frac{1}{2\pi} \exp\left(-\frac{\theta^2}{2}\right)$$

Then, if we observe  $X = 1$ ,

$$\begin{aligned} P(\theta|X = 1) &= \frac{P(X = 1|\theta)P(\theta)}{P(X = 1)} \\ &\propto P(X = 1|\theta)P(\theta) \\ &= \left[ \frac{1}{2\pi} \exp\left(-\frac{(1 - \theta)^2}{2}\right) \right] \left[ \frac{1}{2\pi} \exp\left(-\frac{\theta^2}{2}\right) \right] \\ &\propto N(0.5, 0.5) \end{aligned}$$

*Important Question:* How do we select our prior distribution?

Different possible approaches:

- based on actual prior knowledge about the system or data generation mechanism
- target analytical and computational tractability; e.g., use conjugate priors (those which yield posterior distributions in the same family)
- allow the data to have "maximal impact" on the posterior

## Statistics: Parametric vs. Non-Parametric Models

- All of the models considered above are *parametric* models, in that they are determined by a fixed, finite number of parameters. This can limit the flexibility of the model.
- In contrast, we can endow our models with a potentially infinite number of parameters which is allowed to grow as we see more data. Such models are called *non-parametric*.
- Although non-parametric models yield greater modeling flexibility, they are generally statistically and computationally less efficient.

## Statistics: Generative vs. Discriminative Models

- Suppose that, based on data  $(x_1, y_1), \dots, (x_n, y_n)$ , we would like to obtain a model whereby we can predict the value of  $Y$  based on an always-observed random variable  $X$ .
- *Generative Approach*: model the full joint distribution  $P(X, Y)$ , which fully characterizes the relationship between the random variables.
- *Discriminative Approach*: only model the conditional distribution  $P(Y|X)$
- Both approaches have strengths and weaknesses and are useful in different contexts.

## Matrix Transpose

- For an  $m \times n$  matrix  $A$  with  $(A)_{ij} = a_{ij}$ , its transpose  $A^T$  is defined by  $(A^T)_{ij} = a_{ji}$ .
- $(AB)^T = B^T A^T$

## Matrix Inverse

- The inverse of a square matrix  $A \in \mathbb{R}^{n \times n}$  is the matrix  $A^{-1}$  such that  $A^{-1}A = I$ .
- This notion generalizes to non-square matrices via left- and right-inverses.
- Not all matrices have inverses.
- If  $A$  and  $B$  are invertible, then  $(AB)^{-1} = B^{-1}A^{-1}$ .
- Computation of inverses generally requires  $O(n^3)$  time. However, given a matrix  $A$  and a vector  $b$ , we can compute a vector  $x$  such that  $Ax = b$  in  $O(n^2)$  time.

## Trace

- For a square matrix  $A \in \mathbb{R}^{n \times n}$ , its trace is defined as  $\text{tr}(A) = \sum_{i=1}^n (A)_{ii}$ .
- $\text{tr}(AB) = \text{tr}(BA)$

## Eigenvectors and Eigenvalues

- Given a matrix  $A \in \mathbb{R}^{n \times n}$ ,  $u \in \mathbb{R}^n \setminus \{0\}$  is called an eigenvector of  $A$  with  $\lambda \in \mathbb{R}$  the corresponding eigenvalue if

$$Au = \lambda u$$

- An  $n \times n$  matrix can have no more than  $n$  distinct eigenvector/eigenvalue pairs.

## More definitions

- A matrix  $A$  is called *symmetric* if it is square and  $(A)_{ij} = (A)_{ji}, \forall i, j$ .
- A symmetric matrix  $A$  is *positive semi-definite (PSD)* if all of its eigenvalues are greater than or equal to 0.
- Changing the above inequality to  $>$ ,  $\leq$ , or  $<$  yields the definitions of positive definite, negative semi-definite, and negative definite matrices, respectively.
- A positive definite matrix is guaranteed to have an inverse.

# Linear Algebra: Matrix Decompositions

## Eigenvalue Decomposition

Any symmetric matrix  $A \in \mathbb{R}^{n \times n}$  can be decomposed as follows:

$$A = U\Lambda U^T$$

where  $\Lambda$  is a diagonal matrix with the eigenvalues of  $A$  on its diagonal,  $U$  has the corresponding eigenvectors of  $A$  as its columns, and  $UU^T = I$ .

## Singular Value Decomposition

Any matrix  $A \in \mathbb{R}^{m \times n}$  can be decomposed as follows:

$$A = U\Sigma V^T$$

where  $UU^T = VV^T = I$  and  $\Sigma$  is diagonal.

*Other Decompositions:* LU (into lower and upper triangular matrices); QR; Cholesky (only for PSD matrices)

# Optimization: Basics

- We often seek to find optima (minima or maxima) of some real-valued vector function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ . For example, we might have  $f(x) = x^T x$ .
- Furthermore, we often constrain the value of  $x$  in some way: for example, we might require that  $x \geq 0$ .
- In standard notation, we write

$$\begin{aligned} \min_{x \in \mathcal{X}} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, i = 1, \dots, N \\ & h_i(x) = 0, i = 1, \dots, M \end{aligned}$$

- Every such problem has a (frequently useful) corresponding Lagrange dual problem which lower-bounds the original, primal problem and, under certain conditions, has the same solution.
- It is only possible to solve these optimization problems analytically in special cases, though we can often find solutions numerically.

# Optimization: A Simple Example

- Consider the following unconstrained optimization problem:

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 = \min_{x \in \mathbb{R}^n} (Ax - b)^T (Ax - b)$$

- In fact, this is the optimization problem that we must solve to perform least-squares regression.
- To solve it, we can simply set the gradient of the objective function equal to 0.
- The gradient of a function  $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$  is the vector of partial derivatives with respect to the components of  $x$ :

$$\nabla_x f(x) = \left( \frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} \right)$$

## Optimization: A Simple Example

Thus, we have

$$\begin{aligned}\nabla_x \|Ax - b\|_2^2 &= \nabla_x \left[ (Ax - b)^T (Ax - b) \right] \\ &= \nabla_x \left[ x^T A^T A x - 2x^T A^T b + b^T b \right] \\ &= 2A^T A x - 2A^T b \\ &= 0\end{aligned}$$

and so the solution is

$$x = (A^T A)^{-1} A^T b$$

(if  $(A^T A)^{-1}$  exists).

# Optimization: Convexity

- In the previous example, we were guaranteed to obtain a global minimum because the objective function was *convex*.
- A differentiable function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is convex if its Hessian (matrix of second derivatives) is everywhere PSD (if  $n = 1$ , then this corresponds to the second derivative being everywhere non-negative)<sup>2</sup>.
- An optimization problem is called convex if its objective function  $f$  and inequality constraint functions  $g_1, \dots, g_N$  are all convex, and its equality constraint functions  $h_1, \dots, h_M$  are linear.
- For a convex problem, all minima are in fact global minima. In practice, we can efficiently compute minima for problems in a number of large, useful classes of convex problems.

---

<sup>2</sup>This definition is in fact a special case of the general definition for arbitrary vector functions.