Machine Learning, Spring 2019
EN 2105
MWF 9-9:50am

Chao Lan
EN 4084A
clan@uwyo.edu

Mathematics Review for Machine Learning

Probability

Consider a single random variable Z (e.g. your GPA after completing this semester). If Z takes values from a discrete domain (e.g., $Z \in \{A, B, C, F\}$), it is a discrete variable. If Z takes values from a continuous domain (e.g. $Z \in [0, 4]$), it is a continuous variable.

The randomness of Z can be described using probabilities. The cumulative distribution function (CDF) of Z tells us how likely Z will fall below a threshold z, i.e.,

$$\mathrm{CDF}_Z(z) = \Pr(Z \leq z) =: P(z). \tag{1}$$

If Z is discrete, we can define its probability mass function (PMF) which tells us how likely Z will take a particular value z, i.e.,

$$\mathrm{PMF}_Z(z) = \Pr(Z = z) =: p(z). \tag{2}$$

This implies $P(z) = \sum_{a \leq z} p(a)$.

If Z is continuous and its CDF is differentiable, we can take derivative of its CDF with respect to z and get the probability density function (PDF) of Z[1], i.e.,

$$\mathrm{PDF}_Z(z) = \frac{\partial}{\partial z}\mathrm{CDF}_Z(z) =: p(z). \tag{3}$$

This implies $P(z) = \int_{a \in (-\infty, z)} p(a)$.

There are many well-studied probability distributions, characterized by different PDF or PMF functions with different parameters (which can be estimated from data). For continuous Z, a common distribution is Gaussian or normal distribution; its PDF is

$$p(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(z-\mu)^2}{2\sigma^2}\right\}, \tag{4}$$

where μ and σ^2 are two parameters known as the mean and variance of the distribution, and σ is the standard deviation; if Z is a multivariate variable of dimension p, its PDF is

$$p(z) = \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} \exp\left\{-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right\}. \tag{5}$$

Another common distribution of continuous variable is uniform: if $z \in [a, b]$, its PDF is

$$p(z) = \frac{1}{b-a}. \tag{6}$$

[1] We can interpret PDF as how likely Z will fall in an arbitrarily small interval containing z.

For discrete Z, a common distribution is Bernoulli: if Z is binary, its PMF is

$$p(Z=1)=\theta \quad \text{and} \quad P(Z=0)=(1-\theta), \tag{7}$$

where $\theta \in [0,1]$ is a parameter. We often write Bernoulli distribution as

$$p(z)=\theta^z(1-\theta)^{1-z}. \tag{8}$$

Another common discrete distribution is uniform: if $Z \in \{z_1, \ldots, z_m\}$, its PMF is

$$p(Z=z_i)=\frac{1}{m}. \tag{9}$$

Now consider two random variables Z_1 and Z_2. Their joint CDF is[2]

$$\mathrm{CDF}_{Z_1,Z_2}(z_1,z_2)=\Pr(Z_1 \le z_1 \wedge Z_2 \le z_2)=P(z_1,z_2) \tag{10}$$

If both variables are continuous, their joint PDF is

$$\mathrm{PDF}_{Z_1,Z_2}(z_1,z_2)=\frac{\partial}{\partial z_1 \partial z_2}\mathrm{CDF}_{Z_1,Z_2}(z_1,z_2)=p(z_1,z_2). \tag{11}$$

If both variables are discrete, their joint PMF is

$$\mathrm{PMF}_{Z_1,Z_2}(z_1,z_2)=\Pr(Z_1=z_1 \wedge Z_2=z_2)=p(z_1,z_2). \tag{12}$$

Since machine learning typically works with differentiable CDFs, from now on we always assume PDF exists and discuss probability distributions based on PDF or PMF. In addition, we always assume variables are exchangeable, i.e.

$$p(z_1,z_2)=p(z_2,z_1). \tag{13}$$

The conditional PDF of Z_1 (conditioned on Z_2) is defined as

$$p(z_1 \mid z_2):=\frac{p(z_1,z_2)}{p(z_2)}. \tag{14}$$

[*Discussion*] What is the conditional PDF of Z_2 (conditioned on Z_1)?

Now we have the Bayes' Rule

$$p(z_2 \mid z_1)=\frac{p(z_1 \mid z_2)\cdot p(z_2)}{p(z_1)}. \tag{15}$$

[*Exercise*] Prove the Bayes' Rule.

If there are more than two variables, e.g., Z_1, Z_2 and Z_3, we have the Chain Rule

$$p(z_1,z_2,z_3)=p(z_1)\cdot p(z_2 \mid z_1)\cdot p(z_3 \mid z_1,z_2). \tag{16}$$

[*Exercise*] Prove the Chain Rule.

[2] We will denote logic 'and' by '$\wedge$', and logic 'or' by '$\vee$'.

By the definition of conditional probability, we also have the property

$$p(z_3 \mid z_2, z_1) \cdot p(z_2 \mid z_1) = p(z_3, z_2 \mid z_1). \tag{17}$$

[*Exercise*] Prove (17).

Finally, if D_2 is the domain of z_2, we have

$$p(z_1) = \int_{z_2 \in D_2} p(z_1, z_2) \qquad or \qquad p(z_1) = \sum\nolimits_{z_2 \in D_2} p(z_1, z_2). \tag{18}$$

In (18), the right-to-left process is marginalization, which infers the marginal distribution $p(z_1)$ from joint distribution $p(z_1, z_2)$ by integrating out the randomness of Z_2; the left-to-right process is decomposition, which introduces additional variables (e.g., for easier modeling or estimation).

We say Z_1 and Z_2 are independent if

$$p(z_1 \mid z_2) = p(z_1) \qquad \text{or equivalently} \qquad p(z_1, z_2) = p(z_1)p(z_2), \tag{19}$$

and they are identical if

$$p(z_1) = p(z_2), \tag{20}$$

and they are i.i.d. (independently and identically distributed) if both.

Statistics

The expectation of Z measures its expected value averaged over all possible values, i.e.,

$$E[Z] = \int_z z \cdot p(z) \qquad \text{or} \qquad E[Z] = \sum_z z \cdot p(z). \tag{21}$$

The variance of Z measures its variation around its expectation, i.e.,

$$Var[Z] = \int_z (z - E[Z])^2 \cdot p(z) \qquad \text{or} \qquad Var[Z] = \sum_z (z - E[Z])^2 \cdot p(z). \tag{22}$$

The above definition is often written as

$$Var[Z] = E[(Z - E[Z])^2]. \tag{23}$$

Let Z_1 and Z_2 be two random variables. Their covariance is

$$Cov[Z_1, Z_2] = E[\,(Z_1 - E[Z_1])\,(Z_2 - E[Z_2])\,], \tag{24}$$

where the expectation is taken over the joint randomness of Z_1 and Z_2.

From now on, we will always use Greek letters (e.g., α and β) to denote variables in $\mathbb{R}$.

The following properties are commonly used.

- $E[Z_1 + Z_2] = E[Z_1] + E[Z_2]$

- $E[\alpha Z] = \alpha\, E[Z]$

- $Var[Z_1 + Z_2] = Var[Z_1] + Var[Z_2] + 2Cov[Z_1, Z_2]$

- $Var[\alpha Z] = \alpha^2 Var[Z]$

- If C is a constant, then $Var[C] = 0$.

- $Var[Z] = E[Z^2] - E^2[Z]$

- $Cov[Z_1, Z_2] = E[Z_1 Z_2] - E[Z_1]E[Z_2]$.

- If Z_1, Z_2 are independent, then $Cov[Z_1, Z_2] = 0$.

[*Exercise*] Prove the above properties.

Linear Algebra

Given two d-dimensional vectors $x_1 = [x_{11}, x_{12}, \ldots, x_{1d}]^T$ and $x_2 = [x_{21}, x_{22}, \ldots, x_{2d}]^T$. We have the properties $x_1 + x_2 = [x_{11} + x_{21}, x_{12} + x_{22}, \ldots, x_{1d} + x_{2d}]^T$ and $\alpha x_1 = [\alpha x_{11}, \alpha x_{12}, \ldots, \alpha x_{1d}]^T$.

The inner product of x_1 and x_2 is a scalar, i.e.,

$$\langle x_1, x_2 \rangle = x_1^T x_2 = x_{11}x_{21} + x_{12}x_{22} + \ldots + x_{1d}x_{2d} = \sum\nolimits_{j=1}^{p} x_{1j}x_{2j}. \tag{25}$$

We say x_1 and x_2 are orthogonal if $\langle x_1, x_2 \rangle = 0$.

The norm of vector x_i is a function of x_i measuring its 'length'. There are many measurements, and a general one is Lp-norm, defined as

$$||x_i||_p = \left(\sum\nolimits_{j=1}^{d} (x_{ij})^p\right)^{1/p}. \tag{26}$$

Two common choices are $p = 2$ and $p = 1$; the former gives L2-norm and the latter L1-norm:

$$||x_i||_2 = \sqrt{x_{i1}^2 + x_{i2}^2 + \ldots + x_{id}^2} \qquad \text{and} \qquad ||x_i||_1 = |x_{i1}| + |x_{i2}| + \ldots + |x_{id}|. \tag{27}$$

By definition we have

$$||x_i||_2^2 = \langle x_i, x_i \rangle\,. \tag{28}$$

[*Exercise*] Verify (28).

[*Discussion*] What is the geometric interpretation of vector norm?

Another common concept is L0-norm. It is not a norm by definition, and it counts the number of non-zero elements in a vector, e.g., if $x = [0, 2, 0, 4, 5]$, then $||x||_0 = 3$.

Finally, the outer product of x_1 and any q-dimensional vector x_3 is a p-by-q matrix, i.e.,

$$x_1 x_3^T = \begin{bmatrix} x_{11}x_{31} & x_{11}x_{32} & \cdots & x_{11}x_{3q} \\ x_{12}x_{31} & x_{12}x_{32} & \cdots & x_{22}x_{3q} \\ \vdots & & & \vdots \\ x_{1p}x_{31} & x_{1p}x_{32} & \cdots & x_{1p}x_{3q} \end{bmatrix} \tag{29}$$

Let $U=\{x_1, x_2\}$ be a set of vectors.[3] We say a vector $z \in \mathbb{R}^p$ is linearly dependent on U if it can be linearly expressed by elements in U, i.e., there exists some α_1, α_2 such that

$$z = \alpha_1 x_1 + \alpha_2 x_2. \tag{30}$$

We say z is linearly independent to U if it cannot be linearly expressed by U. We say a set of vectors are linearly independent if each of them is linearly independent to the rest.

[*Discussion*] What is the geometric interpretation of linear independence?

The span of U is the set of vectors that can be linearly expressed by U, i.e.,

$$\text{span}(U) = \{\alpha_1 x_1 + \alpha_2 x_2; \forall \alpha_1, \alpha_2\} \subseteq \mathbb{R}^p. \tag{31}$$

Further, we say U is a basis of $\mathbb{R}^p$ if its elements are linearly independent. For example, U is a span of $\mathbb{R}^2$ if $x_1 = [0, 1]^T$ and $x_2 = [1, 0]^T$.

Let $M \in \mathbb{R}^{p \times q}$ be a matrix. It is diagonal if only diagonal entries M_{ii} are non-zero. It is square if $p = q$. A square matrix M is orthogonal (or, unitary) if $M^T M = M M^T = I$, where I is an identity matrix which has 1 on all diagonal entries and 0 on the rest. A square matrix M is symmetric if $M_{ij} = M_{ji}$ for all indices i, j. The trace of matrix M is the sum of its diagonal elements, i.e., $tr(M) = \sum_{i=1}^{\min\{p,q\}} M_{ii}$.

The row rank of M is the max number of its linearly independent rows, and column rank is the max number of its linearly independent columns. It is proved row rank = column rank, so we often call them the rank of matrix M. We say M has full row rank if its row rank is p (all rows are linearly independent), and has full column rank if its column rank is q (all columns are linearly independent). If M is square, we just say it is full rank if its rank is p. If a matrix does not have full rank, we say it is rank deficient.

Only full-rank square matrix[4] M has an inverse matrix M^{-1}, which is defined as

$$M^{-1} M = I. \tag{32}$$

The row space of M is the span of its row vectors, and the column space (or range space) is the span of its column vectors. It is more often to work with range space, and express it as

$$\text{Range}(M) = \{Mu; \forall u \in \mathbb{R}^q\}. \tag{33}$$

If M is square and there is a pair $(u \in \mathbb{R}^p, \lambda)$ satisfying

$$Mu = \lambda u, \tag{34}$$

then we say u is an eigenvector of M and λ is the associated eigenvalue. A matrix has multiple eigenvectors and eigenvalues, satisfying

$$M = U \Sigma U^T, \tag{35}$$

where $U = [u_1, \ldots, u_p] \in \mathbb{R}^{p \times p}$ with $u_i \in \mathbb{R}^p$ being the i_{th} eigenvector and $\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^{p \times p}$ with λ_i being the eigenvalue of u_i. If M is symmetric, then its eigenvectors are orthogonal.

[3] For convenience we assume two vectors here. But all discussions apply to arbitrary number of vectors.

[4] For rank-deficient or non-squared matrix, we use pseudo-inverse.

Another useful technique is Singular Vector Decomposition (SVD), which can be applied on matrix of arbitrary size. For matrix $M \in \mathbb{R}^{p \times q}$, SVD takes the form

$$M = U \Sigma V^T, \tag{36}$$

where $U \in \mathbb{U}^{p \times p}$ and $V \in \mathbb{U}^{q \times q}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{p \times q}$ is a diagonal matrix. Columns of U are left singular vectors, columns of V are right singular vectors and diagonal entries of Σ are singular values. The number of non-zero singular values equals matrix rank.

The norm of a matrix measures its 'energy'. A common choice is Frobenius norm (or, F-norm):

$$||M||_F = \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{q} (M_{ij})^2}. \tag{37}$$

Consider a function

$$f(x, y) = x^2 + 2y^2 - 3xy + 4. \tag{38}$$

Its derivative with respect to (w.r.t.) a single variable x is

$$\frac{\partial}{\partial x} f = 2x - 3y. \tag{39}$$

Its derivative w.r.t. a vector of variables $z = [x, y]^T$ is

$$\frac{\partial}{\partial z} f = \begin{bmatrix} \frac{\partial}{\partial x} f \\ \\ \frac{\partial}{\partial y} f \end{bmatrix} = \begin{bmatrix} 2x - 3y \\ \\ 4y - 3x \end{bmatrix}. \tag{40}$$

Its second-order derivative w.r.t. z is its Hessian matrix

$$\frac{\partial^2}{\partial z^2} f = \begin{bmatrix} \frac{\partial}{\partial x \partial x} f & \frac{\partial}{\partial x \partial y} f \\ \\ \frac{\partial}{\partial y \partial x} f & \frac{\partial}{\partial y \partial y} f \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ \\ -3 & 4 \end{bmatrix}. \tag{41}$$

More generally, let $F = [f_1, f_2, \ldots, f_n]$ be a vector of functions depending on variables $x_1, \ldots, x_p$, and $X = [x_1, x_2, \ldots, x_p]^T$ be the vector of variables. The derivative of F w.r.t. X is

$$\frac{\partial F}{\partial X} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_1} \\ \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_2} \\ \\ \vdots & & & \\ \frac{\partial f_1}{\partial x_p} & \frac{\partial f_2}{\partial x_p} & \cdots & \frac{\partial f_n}{\partial x_p} \end{bmatrix} \tag{42}$$

Let $M \in \mathbb{R}^{p \times q}$, $S \in \mathbb{R}^{p \times p}$, $x \in \mathbb{R}^p$, $y \in \mathbb{R}^q$ and $z \in \mathbb{R}^p$. The following derivatives are commonly used in machine learning. (See "The Matrix Cookbook" by the Petersen's for more.)

- $\frac{\partial}{\partial x} x^T z = z$

- $\frac{\partial}{\partial x} x^T M = M$

- $\frac{\partial}{\partial x} M y = M^T$

- $\frac{\partial}{\partial x} x^T S x = (S + S^T)x$

- if S is symmetric, $\frac{\partial}{\partial x} x^T S x = 2Sx$

Optimization

Let $L : F \to \mathbb{R}$ be a function. We will review the optimization problem of finding a point $f \in F$ that minimizes $L(f)$ – the minimization problem. [5]

Let $f_* \in F$. We say $f_* \in F$ is a global minimum point of L if $L(f_*) \leq L(f)$ for any $f \in F$, and the corresponding $L(f_*)$ is a global minimum of L.

We say $f_* \in F$ is a local minimum point of L if $L(f_*) \leq L(f)$ for any $f \in N(f_*)$ where $N(f_*)$ is some neighborhood of f_*, and the corresponding $L(f_*)$ is a local minimum of L.

[*Discussion*] What are the geometric interpretations of global/local minimum point?

We say f_* is a critical point if $L'(f_*) = 0$. A critical point is either a local minimum point, a local maximum point or a saddle point.[6] To verify if a critical point f_* is local maximum or minimum, we can use the second derivative test. Assume L is twice-differentiable at f_*. If $L''(f_*) > 0$ then f_* is a local minimum; if $L''(f_*) < 0$ then f_* is a local maximum; if $L''(f_*) = 0$ then the test is not informative.

A convex combination of two elements $f_1, f_2 \in F$ is

$$\alpha_1 f_1 + \alpha_2 f_2 \tag{43}$$

for some $\alpha_1, \alpha_2 \in [0, 1]$ satisfying $\alpha_1 + \alpha_2 = 1$.

We say F is a convex set if all convex combinations of its any two elements remain in F.

[*Discussion*] What is the geometric interpretation of convex set?

If F is convex, we say L is a convex function if for any $f_1, f_2 \in F$ and $\alpha_1, \alpha_2 \in [0, 1]$ satisfying $\alpha_1 + \alpha_2 = 1$, there is

$$L(\alpha_1 f_1 + \alpha_2 f_2) \leq \alpha_1 L(f_1) + \alpha_2 L(f_2). \tag{44}$$

We say L is a concave function if

$$L(\alpha_1 f_1 + \alpha_2 f_2) \geq \alpha_1 L(f_1) + \alpha_2 L(f_2). \tag{45}$$

We will focus on convex function.[7] An important result is

Remark 1. *If L is convex, then its local minimum point is also its global minimum point.*

If $f \in \mathbb{R}$, we say L is a quadratic function of $f \in \mathbb{R}$ if it has the form

$$L(f) = af^2 + bf + c, \quad a \neq 0, \tag{46}$$

[5] Maximizing L is equivalent to minimizing $-L$.

[6] Saddle point is rare in machine learning (except in neural network) so we will not elaborate on it.

[7] A concave function can be transformed to convex by flipping the sign of its output.

and af^2 is the quadratic term. If $f \in \mathbb{R}^p$, its multivariate quadratic function has the form

$$L(f) = f^T Af + B^T f + c, \quad A \in \mathbb{R}^{p\times p}, B \in \mathbb{R}^p, c \in \mathbb{R}. \tag{47}$$

A quadratic function is convex if $a > 0$ or A is positive semidefinite.

It is easier to work with unconstrained optimization problem, which has the form

$$\min_f L(f), \tag{48}$$

where $L(f)$ is the objective function. It is called unconstrained because there is no constraint.

Global minimum points can be hard to find, but local minimum points are much easier and they work just fine in many machine learning tasks. If L is differentiable and $L'(f) = 0$ is easy to solve, we can find its local minimum point using the 'critical point method'[8] – just look for a critical point and it will become a local/global minimum point under proper conditions.

Algorithm 1 Critical Point Method

```
Assumption: $L$ is differentiable and $L'(f) = 0$ is easy to solve.
1: Calculate the derivative of $L(f)$, denoted by $L'(f)$.
2: Set $L'(f) = 0$ and solve for $f$.
3: Denote a solution by $f_*$. It is an analytic solution or closed-form solution.
4: $f_*$ is a local/global minimum point under proper conditions (e.g., $L$ is convex).
```

[*Exercise*] Find a local minimum point of $L(f)$=$5f^2 - 3f + 10$ on $\mathbb{R}$ by the critical point method.

If L is differentiable but $L'(f) = 0$ is not easy to solve, then critical point method is not suitable. In this case, we can try numerical optimization methods that often search for approximate local minimum points in iterative ways. A common one is gradient descent, which iteratively updates solution f_* by the rule

$$f_* = f_* - \eta L'(f_*), \tag{49}$$

where $\eta > 0$ is learning rate. It converges to local minimum under proper conditions.

Algorithm 2 Gradient Descent Method

```
0: (randomly) initialize a solution $f_*$
for t = 0, 1, 2, ..., max_iter do
  1: if stopping criterion is met (e.g. $\Delta L' < \epsilon$), stop and output $f_*$ as the solution
  2: compute gradient $\Delta f = L'(f_*)$
  3: choose a learning rate $\eta$
  4: update solution $f_* = f_* - \eta \cdot \Delta f$
end for
```

[*Discussion*] What is the geometric interpretation of gradient descent method?

[*Exercise*] Find a local minimum point of $L(f) = e^f + 5f^2 - 3f$ using gradient-descent method.

A constrained optimization problem has the form

$$\min_f L(f), \quad s.t.\ g(f) \leq 0. \tag{50}$$

We call $g(f) \leq 0$ a constraint.[9]

[8] The instructor made this name up to facilitate discussion.

[9] There can be many constraints and here we just assume one standard constraint.

A 'lazy' way to solve constrained optimization is to first transform it to unconstrained optimization problem, and then apply the latter's solvers. A common method is the Lagrange multiplier. It augments the constraint to the original objective function, generating the Lagrangian function

$$L_+(f, \lambda) = L(f) + \lambda\, g(f), \tag{51}$$

where λ is the multiplier. Then it solves the unconstrained optimization problem

$$\min\nolimits_{f,\lambda} L_+(f, \lambda). \tag{52}$$

One relation between the original problem and the transformed problem is specified as follows.

Theorem 2. *If f_* is a solution to (50), then there exists a λ_* such that (f_*, λ_*) is a critical point of L_+.*

The theorem specifies a necessary condition for a critical point of L_+ being a solution to (50). But not all critical points of L_+ are solutions to (50). Some sufficient conditions also exist, known as the KKT conditions. We will introduce them when discussing SVM.