Cal Poly **CSC 566** Advanced Data Mining Alexander Dekhtyar

Introduction to Machine Learning.
Part 1: Data

Machine Learning Problems

Machine Learning Problem Description. The most general description of Machine Learning is something like this:

> *Given a number of observations, predict what other observations would look like.*

Machine Learning problems can be classified into

- Regression problems
- Classification problems
- Clustering problems
- Recommendation problems

depending on the shape of the observations, and on the form which the prediction is to take.

Before proceeding with more detailed discussion of different types of machine learning problems, we fisrt must examine the notion of *observations*.

Data for Machine Learning: Objects, Features, Targets

In most cases, data for machine learning problems is a collection of *object descriptions*, where each *object* is described as a *collection of features*.

Features. A *feature* is an individual measurable property of an object or phenomenon that is being observed. A *feature* has the following properties:

- Name. An unique identifier that distinguishes the feature from all other features considered as part of the same machine learning problem.
- Domain. A set of values the feature can take (i.e., the set of possible measurements that can be made of the feature). Feature ranges can be finite or infinite.

Types of features. Depending on the specific range of feature values, features can be classified into the following categories:

1. Numeric. Numeric features have domains that are infinite or finite (but usually rather large) sets of numbers.
2. Categorical. Categorical features are features with domains that are not numeric. Among categorical features, we identify
   (a) Nominal features. These are features whose possible values admit no order.
   (b) Ordinal features. These features have possible values that can be meaningfully ordered, and thus, can be represented as numbers $\{1, \ldots, N\}$ where N is the size of the domain of the feature.

Example. Temperature of a hospital patient patient measured in degrees Celcius is a *numeric feature* whose domain is the range $[30.0, 45.0]$.

The patient's native language is a *categorical nominal feature.* Its domain may be rather large (total number of languages in the world is around 6900), by the possible values of this feature, e.g., ”English”, ”Spanish”, ”Mandarin”, ”Russian”, etc. do not have an ascribed order.

The patient's level of pain is a *categorical ordinal feature*. Its domain may be $\{\mathsf{no}, \mathsf{slight}, \mathsf{moderate}, \mathsf{severe}\}$. There is a natural order that can be imposed on these values: $\mathsf{no} < \mathsf{slight} < \mathsf{moderate} < \mathsf{severe}$, which makes this feature *ordinal*. The domain can be represented as $\{1, 2, 3, 4\}$ (or $\{0, 1, 2, 3\}$) with numbers retaining their property of order.

Synonyms. We use the terms feature, attribute, and variable (or independent variable) interchangebly.

Observations. In *statistics*, **an observation** is a measured value, at a particular moment of time, of a specific feature.

In this class, we use the term observation in a more general way to refer to a collection of measured values of a particular set of features, that *describe a specific object or phenomenon.*

Formalizing. A given machine learning problem describes a set of objects or a set of phenomena by establishing a *set of features* $\mathcal{A} = \{A_1, \ldots, A_n\}$ whose values combined produce a *complete* (from the perspective of the specific machine learning problem) description of a single object/single phenomenon. The values for each feature $A_i, i = 1 \ldots n$ come from the set $D_i = dom(A_i)$.

A vector $\mathbf{x} = (x_1, \ldots, x_n)$ of values, where $(\forall i = 1 \ldots n)(x_i \in dom(A_i))$ is called a **data point**. Often, without loss of generality, we refer to data points as *objects*, *phenomena*, *entities*, *points*, *records*, *tuples*, *feature-vectors* or use a domain-specific name to identify them.

Example. Consider a collection of features $\{\mathsf{Name}, \mathsf{Language}, \mathsf{Temperature}, \mathsf{PainLevel}\}$ representing the name of a hospital patient, their native language, their temperature at admission time, and their self-reported pain level.

A description of a signle patient may be a vector of values

$$\mathbf{x} = (\text{"}\mathsf{Mary\ Smith}\text{"}, \text{"}\mathsf{English}\text{"}, \mathsf{38.2}, \text{"}\mathsf{moderate}\text{"}).$$

We can refer to $\mathbf{x}$ in a number of ways:

- *vector of feature values*, *feature-vector*, or simply *vector*
- *data point*
- *observation*
- *object* (although in this case it is a somewhat awkwards term)
- *record* or *tuple*
- *patient record* (this is a domain-specific term that carries knowledge of the semantics of the data)
- *patient*

Dataset. Let $\mathcal{A} = \{A_1, \ldots, A_n\}$ be a set of features. Let $X = \{\mathbf{x_1}, \ldots, \mathbf{x_m}\}$, where

$$(\forall i = 1 \ldots m)\mathbf{x_i} = (x_{i1}, \ldots, x_{in}),$$

and X is called a *collection of datapoints*, or a *dataset*.

Algebraic View of Datasets

Dataset as a matrix. A dataset $X = \{\mathbf{x_1}, \ldots, \mathbf{x_m}\}$ can be viewed as a *matrix* X constructed as follows:

- The *rows* of X are vectors $\mathbf{x_1}, \ldots \mathbf{x_m}$.
- The *columns* of X are individual features $A_1, \ldots, A_n$, with each column containing the values of a single feature from all data points of the dataset.

Dataset X can be visualized as a matrix as follows:

$$X = \left(\begin{array}{c|cccc} & A_1 & A_2 & \dots & A_n \\ \hline \mathbf{x_1} & x_{11} & x_{12} & \dots & x_{1n} \\ \mathbf{x_2} & x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x_m} & x_{m1} & x_{m2} & \dots & x_{mn} \end{array} \right)$$

Size of a dataset. The **size** of a dataset $X = \{\mathbf{x_1}, \dots \mathbf{x_n}\}$, denoted $|X|$ is the number of data points in it, i.e.,

$$|X| = m$$

The **dimensionality** of the dataset X, denoted $dim(X)$ is the length of the vectors in it. If

$$\mathbf{x_i} = (x_{i1}, \dots x_{in}),$$

then

$$dim(X) = n.$$

Numeric vector-spaces. Let $\mathcal{A} = \{A_1, \dots, A_n\}$ and let $(\forall i \in 1 \dots n) dom(A_i) \subseteq \mathbb{R}$. In this case, given a dataset $X = \{\mathbf{x_1}, \dots, \mathbf{x_n}\}$ of points in the feature-space $\mathcal{A}$, we can write for each data point $\mathbf{x_i}$ that

$$\mathbf{x_i} = (x_{i1}, \dots, x_{in}) \in \mathbb{R}^n,$$

i.e., $\mathbf{x_i}$ is a point in n-dimensional real space.

Alternatively, we can view $\mathbf{x_i}$ as an n-dimensional vector (vectors in real space are considered to be columns):

$$\mathbf{x_i} = \left(\begin{array}{c} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{array} \right) = (x_{i1}, \dots, x_{in})^T.$$

Linear Representation of vectors. Let

$$I = \left(\begin{array}{cccc} \mathbf{e_1} & \mathbf{e_2} & \dots & \mathbf{e_n} \\ \hline 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{array} \right)$$

be a matrix of unit vectors

$$\mathbf{e_i} = (0, 0, \dots 1_i, 0, \dots, 0)^T.$$

The vectors $\mathbf{e_1}, \dots \mathbf{e_n}$ are said to form the *standard basis* of $\mathbb{R}^n$.

Vector $\mathbf{x_i}$ can be represented as a linear combination of $e_1, \dots, e_n$ as

$$\mathbf{x_i} = x_{i1}\mathbf{e_1} + x_{i2}\mathbf{e_2} + \dots + x_{in}\mathbf{e_n} = \sum_{j=1}^{n} x_{ij}\mathbf{e_j}.$$

More notation. From the above, we can denote/represent the dataset $X = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ as follows:

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \ldots & x_{mn} \end{pmatrix} = \begin{pmatrix} -\mathbf{x_1}^T- \\ -\mathbf{x_2}^T- \\ \vdots \\ -\mathbf{x_m}^T- \end{pmatrix} = \begin{pmatrix} | & | & & | \\ A_1 & A_2 & \ldots & A_n \\ | & | & & | \end{pmatrix}$$

Data Point Manipulations

Let $X = \{\mathbf{x_1}, \ldots, \mathbf{x_m}\} \subseteq \mathbb{R}^n$ be a dataset of numeric data points, and let $\mathbf{x} \in X$ and $\mathbf{y} \in X$ be two data points from X:

$$\mathbf{x} = (x_1 \ldots, x_n)^T$$
$$\mathbf{y} = (y_1 \ldots, y_n)^T$$

We recall a number of important operations on individual n-dimensional vectors, and pairs of such vectors. This operations *play an important role in machine learning methodology.*

Dot product. The dot product of $\mathbf{x}$ and $\mathbf{y}$, denoted $\mathbf{x} \cdot \mathbf{y}$ is defined as

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T\mathbf{y} = (x_1, \ldots, x_n) \times \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} =$$
$$= x_1y_1 + x_2y_2 + \ldots x_ny_n = \sum_{i=1}^{n} x_iy_i.$$

Orthogonality. Two vectors $\mathbf{x}$ and $\mathbf{y}$ are called **orthogonal** iff

$$\mathbf{x} \cdot \mathbf{y} = 0.$$

Vector norms. The L_1-*norm* of a vector $\mathbf{x}$ is defined as

$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + \ldots + |x_n| = \sum_{i=1}^{n} |x_i|.$$

The L_2-*norm* of vector $\mathbf{x}$, also called the *Eucledean norm* or the *length* of $\mathbf{x}$ is defined as

$$\|\mathbf{x}\|_2 = \|\mathbf{x}\| = \sqrt{\mathbf{x}^T \cdot \mathbf{x}} = \sqrt{x_1^2 + x_2^2 + \ldots x_n^2} = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

In general, an L_p-*norm* of a vector $\mathbf{x}$ is defined as

$$\|\mathbf{x}\|_p = (x_1^p + x_2^p + \ldots + x_n^p)^{\frac{1}{p}} = \left(\sum_{i=1}^{n} x_i^p\right)^{\frac{1}{p}}.$$

Distance. The *Eucledean distance* between two vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\delta(\mathbf{x},\mathbf{y}) = \|x - y\| = \sqrt{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y}} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}.$$

Similarly, the L_p-norm distance between $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\delta_p(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_p$$

In particular, the L_1-norm distance between $\mathbf{x}$ and $\mathbf{y}$, also known as the *Manhattan distance* is

$$Manhattan(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_1 = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_n - y_n| = \sum_{i=1}^{n}|x_i - y_i|.$$

Angle. In many settings, it is helpful to compare the *directions* of the n-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$ from the dataset X. In such cases, we use the cosine similarity score, which is the cosine of the angle between the two vectors. It is defined as follows:

$$\cos(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = \left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)^T\left(\frac{\mathbf{y}}{\|\mathbf{y}\|}\right) =$$
$$= \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}.$$

Cauchy-Scwartz inequality. The *Cauchy-Schwarts inequality* for the L_2-norm and the dot product of two vectors state that

$$|\mathbf{x}\cdot\mathbf{y}| \leq \|\mathbf{x}\|\|\mathbf{y}\|$$

Therefore,

$$-1 \leq cos(\mathbf{x},\mathbf{y}) \leq 1$$

Mean point of a dataset. Given a dataset $X = \{\mathbf{x_1},\ldots,\mathbf{x_m}\}$, the mean point of the dataset or the mean is defined as:

$$mean(X) = \mu_{\mathbf{X}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x_i}$$

Total Variance. The total variance of the dataset X is *the average squared distance from a point in X to the mean point of X*:

$$var(X) = \frac{1}{m}\sum_{i=1}^{m}\delta(\mathbf{x_i},\mu_{\mathbf{X}}) = \frac{1}{m}\sum_{i=1}^{m}\|\mathbf{x_i}-\mu_{\mathbf{X}}\|^2$$

We can simplify the computation of variance as follows:

$$var(X) = \frac{1}{m}\sum_{i=1}^{m}\|\mathbf{x_i} - \mu_{\mathbf{X}}\|^2 = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \mu_{Xj})^2 =$$

$$\frac{1}{m}\sum_{i=1}^{m}(\|\mathbf{x_i}\|^2 - 2\mathbf{x_i}^T\mu_{\mathbf{X}} + \|\mu_{\mathbf{X}}\|^2 = \frac{1}{m}\left(\sum_{i=1}^{m}\|\mathbf{x_i}\|^2 - 2m\mu_{\mathbf{X}}^T\left(\frac{1}{m}\sum_{i=1}^{m}\mathbf{x_i}\right) + m\|\mu_{\mathbf{X}}\|^2\right) =$$

$$\frac{1}{m}\left(\sum_{i=1}^{m}\|\mathbf{x_i}\|^2 - 2m\mu_{\mathbf{X}}^T\mu_{\mathbf{X}} + m\|\mu_{\mathbf{X}}\|^2\right) =$$

$$= \frac{1}{m}\left(\sum_{i=1}^{m}\|\mathbf{x_i}\|^2\right) - \|\mu_{\mathbf{X}}\|^2$$

Orthogonal Projection. Consider two vectors (data points) $\mathbf{x}, \mathbf{y} \in X$.

An **orthogonal decomposition** of $\mathbf{x}$ in the direction of $\mathbf{y}$ is a pair of vectors $\mathbf{p}$, $\mathbf{r}$, such that the following holds:

$$\mathbf{x} = \mathbf{p} + \mathbf{r}$$
$$\mathbf{p}^T\mathbf{r} = 0$$
$$\mathbf{r}^T\mathbf{y} = 0$$

That is, $\mathbf{x}$ can be decomposed as the sum of two *orthogonal vectors*, $\mathbf{p}$ and $\mathbf{r}$, one of which ($\mathbf{r}$) is, in turn, *orthogonal to* $\mathbf{y}$ (which forces the other vector, $\mathbf{p}$ to be co-aligned, or *colinear* with $\mathbf{y}$)

Vector $\mathbf{p}$ is the **orthogonal projection** of $\mathbf{x}$ onto $\mathbf{y}$. Vector $\mathbf{r} = \mathbf{x} - \mathbf{p}$ is the **perpendicular distance** between $\mathbf{x}$ and $\mathbf{y}$, and is colinear with the *normal vector* to $\mathbf{y}$.

Because $\mathbf{p}$ is colinear with $\mathbf{y}$, we have

$$\mathbf{p} = c\mathbf{y},$$

and

$$\mathbf{r} = \mathbf{x} - c\mathbf{y}.$$

Because $\mathbf{p}^T\mathbf{r} = 0$, we get

$$\mathbf{p}^T\mathbf{r} = (c\mathbf{y})^T(\mathbf{x} - c\mathbf{y}) = c\mathbf{y}^T\mathbf{x} - c^2\mathbf{y}^T\mathbf{y} = 0$$

Because $c \neq 0$, we must therefore have

$$\mathbf{y}^T\mathbf{x} - c\mathbf{y}^T\mathbf{y} = 0,$$

which yields

$$c = \frac{\mathbf{y}^T\mathbf{x}}{\mathbf{y}^T\mathbf{y}} = \frac{\mathbf{y}\cdot\mathbf{x}}{\|\mathbf{y}\|^2}$$

Therefore, the orthogonal projection of $\mathbf{x}$ onto $\mathbf{y}$ can be computed as

$$\mathbf{p} = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{y}\|^2}\mathbf{x}.$$

Another form of this expression is

$$\mathbf{p} = \left(\|\mathbf{x}\|^2 \cos(\mathbf{x}, \mathbf{y})\right) \mathbf{x}.$$

More Linear Algebra Concepts

Linear combinations. Let $X = \{\mathbf{x_1}, \ldots, \mathbf{x_m}\}$ be a dataset of points $\mathbf{x_i} \in \mathbb{R}^n$. Given a set of scalar values $c_1, \ldots, c_m$, the vector

$$c_1\mathbf{x_1} + c_2\mathbf{x_2} + \ldots + c_m\mathbf{x_m}$$

is called a *linear combination* of vectors $\mathbf{x_1}, \ldots, \mathbf{x_m}$.

Spanning set. The set of all possible linear combinations

$$\mathbf{v} = \sum_{i=1}^{m} c_i\mathbf{x_i}$$

of vectors $\mathbf{x_1}, \ldots, \mathbf{x_m}$, denoted $span(\mathbf{x_1}, \ldots, \mathbf{x_m})$ or $span(X)$ is called the **spanning set** of $\mathbf{x_1}, \ldots, \mathbf{x_m}$.

Row and Column Space. Given the dataset X represented as an $m \times n$ matrix:

$$X = \left(\begin{array}{c|cccc} & X_1 & X_2 & \ldots & X_n \\ \hline \mathbf{x_1} & x_{11} & x_{12} & \ldots & x_{1n} \\ \mathbf{x_2} & x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x_m} & x_{m1} & x_{m2} & \ldots & x_{mn} \end{array}\right)$$

the **column span** of X, denoted $col(X)$ is defined as the

$$col(X) = span(X_1, \ldots, X_n),$$

and the **row span** of X, denoted $row(X)$ is defined as the

$$row(X) = span(\mathbf{x_1}, \ldots, \mathbf{x_m})$$

We can see that $col(X) \subseteq \mathbb{R}^m$ and $row(X) \subseteq \mathbb{R}^n$.

Linear dependence. A set of vectors $\mathbf{v_1}, \ldots, \mathbf{v_k}$ are *linearly dependent* iff at least one vector can be represented as a linear combination of others, i.e., if there are scalars $c_1, \ldots, c_k$ such that at least one $c_i \neq 0$, and

$$c_1\mathbf{v_1} + \ldots c_k\mathbf{v_k} = 0$$

Vectors $\mathbf{v_1}, \ldots, \mathbf{v_k}$ are *linearly independent* otherwise.

Rank. Let $S \subseteq \mathbb{R}^d$. A *basis set* for S is a set of linearly independent vectors $B = \{\mathbf{v_1}, \ldots, \mathbf{v_k}\}$ such that $span(\mathbf{v_1}, \ldots, \mathbf{v_k}) = S$.

If all vectors in B are pairwise orthogonal,i.e., $(\forall i, j \in 1 \ldots k, i \neq j)(\mathbf{v_i} \cdot \mathbf{v_j} = 0)$, B is called an *orthogonal basis* of S.

If for each vector $v_i \in B$, $\|v_i\| = 1$, *and* B is an orthogonal basis, then, B is called *an orthonormal basis*.

The *standard basis* for $\mathbb{R}^d$ is the basis consisting of vectors

$$\begin{aligned} \mathbf{e_1} &= (1, 0, \ldots, 0)^T \\ \mathbf{e_2} &= (0, 1, \ldots, 0)^T \\ &\ldots \\ \mathbf{e_d} &= (0, 0, \ldots, 1)^T \end{aligned}$$

Theorem. Every basis of a set S of vectors *has the same number of vectors*.

The number of vectors in a basis of S is called the *dimension of* S, denoted $dim(S)$.

Theorem. Given a matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix},$$

$$dim(col(A)) = dim(row(A)).$$

This value is called the *rank* of the matrix A, denoted $rank(A)$.