10-601A: Machine Learning, Spring 2015

Probability/Statistics Review & Linear Regression

Lecturer: Roni Rosenfeld *Scribe: Udbhav Prasad*

1 Probability and Statistics

A regular variable holds, at any one time, a single value, be it numeric or otherwise. In contrast, a *random variable* (RV) holds a *distribution* over values, be they numeric or otherwise. For example, the outcome of a future toss of a fair coin can be captured by a random variable X holding the following distribution: 'HEAD' with probability 0.5, and 'TAIL' with probability 0.5 .

We will use uppercase characters to denote random variables, and their lowercase equivalents to denote the values taken by those random variables. The following are commonly used notations to represent probabilities:

1. $\Pr(x)$ is shorthand for $\Pr(X = x)$
2. $\Pr(x, y)$ is shorthand for $\Pr(X = x \text{ AND } Y = y)$
3. $\Pr(x \mid y)$ is shorthand for $\Pr(X = x \mid Y = y)$

In a multivariate distribution over X and Y, the **marginal** of X is

$$\Pr_X(x) = \sum_y \Pr(x, y)$$

and the marginal of Y is

$$\Pr_Y(y) = \sum_x \Pr(x, y)$$

The **Chain Rule** is:

$$\Pr(x, y) = \Pr(x \mid y) \Pr(y) = \Pr(y \mid x) \Pr(x)$$

Independence of two random variables X and Y is defined as follows ($\perp\!\!\!\perp$ is the symbol for independence):

$$X \perp\!\!\!\perp Y \iff \forall x \in X, y \in Y, \Pr(x, y) = \Pr(x) \Pr(y)$$

$$X \perp\!\!\!\perp Y \iff Y \perp\!\!\!\perp X$$

Expected value or mean of a RV is defined (for a discrete RV) as:

$$\mathrm{E}[X] = \sum_x x \Pr(x)$$

Properties of E:

1. E is a linear operator: $\mathrm{E}[aX + b] = a\mathrm{E}[X] + b$.
2. $\mathrm{E}[aX + bY + c] = a\mathrm{E}[X] + b\mathrm{E}[Y] + c$ (note that this doesn't assume any relationship between X and Y).
3. $\mathrm{E}[\sum_i f_i(X, Y, Z \ldots)] = \sum_i \mathrm{E}[f_i(X, Y, Z \ldots)]$ where f_is are some functions of the random variables. Again, note that this does not assume anything about the f_is or the relationship among $X, Y, Z \ldots$.
4. In general, $\mathrm{E}[f(X)] \neq f(\mathrm{E}[X])$. For example, $\mathrm{E}[X^2]$ is often not equal to $(\mathrm{E}[X])^2$ (incidentally $(\mathrm{E}[X])^2$ can also be denoted as $\mathrm{E}^2[X]$). In fact, for the specific case of $f(X) = X^2$, it is always true that $\mathrm{E}[f(X)] \geq f(\mathrm{E}[X])$.

Variance $\mathrm{Var}[X]$ of a random variable X (also denoted as $\sigma^2(X)$) is defined as: $\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2]$

Variance is the second moment of the RV about its mean (also known as the second central moment). Its units are the square of the units of the original RV.

A useful alternative formula for the variance: $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}^2[X]$

Standard deviation $\sigma(X)$ of X is defined as: $\sigma(X) = \sqrt{\mathrm{Var}[X]}$

The **Covariance** of X and Y, also written as $\sigma(X, Y)$, is defined as:
$\mathrm{Cov}[X, Y] = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] = \mathrm{E}[(XY - \mathrm{E}[X]Y - X\mathrm{E}[Y] + \mathrm{E}[X]\mathrm{E}[Y])] = \mathrm{E}[XY] - (\mathrm{E}[X])(\mathrm{E}[Y])$.
So $\mathrm{Cov}[X, Y] = 0 \iff \mathrm{E}[XY] = (\mathrm{E}X)(\mathrm{E}Y)$.

Properties of Var, σ and Cov (a, b and c are real constants):

1. $\mathrm{Var}[X + b] = \mathrm{Var}[X]$.
2. $\mathrm{Var}[aX] = a^2\mathrm{Var}[X]$.
3. $\mathrm{Var}[aX + b] = a^2\mathrm{Var}[X]$.
4. $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X, Y]$
5. $\mathrm{Var}[aX + bY + c] = \mathrm{Var}[aX + bY] = a^2\mathrm{Var}[X] + b^2\mathrm{Var}[Y] + 2ab\mathrm{Cov}[X, Y]$.
6. $\mathrm{Var}[\sum_i w_i X_i] = \sum_{i,j} w_i w_j \mathrm{Cov}[X_i, X_j]$. (Note that $\mathrm{Var}[X_i] = \mathrm{Cov}[X_i, X_i]$)
7. Variance of uncorrelated variables is additive: $(\forall i, \forall j \neq i \ \mathrm{Cov}[Xi, Xj] = 0) \implies \mathrm{Var}[\sum_i X_i] = \sum_i \mathrm{Var}[X_i]$.
8. $\mathrm{Var}[X] = \mathrm{Cov}[X, X]$.
9. $\sigma(aX + b) = |a| * \sigma(X)$.
10. $\mathrm{Cov}[aX + b, cY + d] = \mathrm{Cov}[aX, cY] = ac\mathrm{Cov}[X, Y]$. Covariance is invariant under shift of either variable.

The Law of Total Variance: $\mathrm{Var}[Y] = \mathrm{Var}_X[\mathrm{E}[Y \,|\, X]] + \mathrm{E}_X[\mathrm{Var}[Y \,|\, X]]$
In clustering, the first term is the *cross-cluster variance*, and the second term is the *within-cluster variance*.
In regression, the first term is the *explained variance*, and the second term is the *unexplained variance*.

Linear Correlation, Corr$[X, Y]$, often loosely called just 'correlation', is defined as

$$\text{Corr}[X, Y] = \rho(X, Y) = \frac{\text{Cov}[X, Y]}{\sigma(X)\sigma(Y)} \quad \in [-1, +1]$$

Correlation is invariant under shift and scale of either variable (up to change of sign):

$$\rho(aX + b, cY + d) = sign(ac)\rho(X, Y)$$

$$X' = (X - \text{E}[X])/\sigma(X)$$
$$Y' = (Y - \text{E}[Y])/\sigma(Y)$$

X', Y' are zero-mean, unit-variance $\implies \rho(X, Y) = \rho(X', Y') = \text{E}[X'Y']$.

$X \perp\!\!\!\perp Y \implies \rho = 0$, but not vice versa! X, Y can be (linearly) uncorrelated, but still dependent! (think of the case of distribution of points on the circumference of a circle). Linear correlation measures only the extent to which X, Y are *linearly* related and not some other relationship between them.

Summary: Independent $\implies$ uncorrelated.

2 Correlation vs. Mutual Information

Recall the definitions of Mutual Information

$$\text{I}(X; Y) = \text{E}\Big[\log \frac{\Pr(x, y)}{\Pr(x)\Pr(y)}\Big]$$

and Linear Correlation

$$\text{Corr}[X, Y] = \rho(X, Y) = \frac{\text{Cov}[X, Y]}{\sigma(X)\sigma(Y)}$$

(Linear) Correlation requires the x, y values to be numerical, so that there is a notion of *distance* between two values — a metric space. It measures the **extent** or **tightness** of linear association, not its slope. Linear correlation is invariant under a linear transformation (shifting and scaling) of either variable on its own. Note however that correlation is *not* invariant to rotation, because rotation of the (X, Y) plane corresponds to a *joint* linear transformation of the two variables: the new X variable is a function of both the old X and the old Y, and similarly for the new Y variable. This allows the correlation to change.

To calculate correlation between two binary RVs, treat each RV as having any two numerical values, say 0 and 1 — it doesn't matter what two values are chosen.

ρ is dimensionless – it is a pure number. Its range: $\rho \in [-1, +1]$. Often we care only about the *strength* of the correlation, rather than its polarity. In that case we tend to look at ρ^2 which is in the range $[0, 1]$. In fact, ρ^2 has an important interpretation: it is the *fraction* of the variance of Y that's explained by X (compare to the Law of Total Variance above).

In contrast, **mutual information** does *not* require a metric space. X and/or Y could take on any values, *e.g.* $X = \{blue, red, green\}$, $Y = \{math, physics\}$.

Range: $0 \leq \text{I}(X, Y) \leq \min(H(X), H(Y))$
Dimension: bits.
$\text{I}(X; Y) = 0 \iff X \perp\!\!\!\perp Y$.

Examples of high mutual information ($\text{I}(X; Y)$) but correlation (ρ) $= 0$:

1. A perfect polygon, with uniform probability distribution on the vertices. $\rho(X,Y)$ is always 0, but as the number of vertices goes to infinity, so does $\mathrm{I}(X;Y)$.
2. Smallest example: Uniform distribution on the vertices of an equilateral triangle.
3. Consider a uniform distribution over the vertices of a square, and consider what happens when you rotate the square. Correlation is preserved (at 0), because rotation corresponds to linear transformation of each random variable. But mutual information is not invariant to rotation! When the square is axis parallel, $I(X;Y)$ is reduced.

$$\mathrm{I}(X;Y) = 0 \iff X \perp\!\!\!\perp Y \implies \rho(X,Y) = 0$$

Can we have zero mutual information but non-zero correlation? No, because $\mathrm{I}(X;Y) = 0$ means X, Y are independent, so $\rho = 0$. But we can have high correlation (1.0) with arbirarily low mutual information. For example, consider the 2x2 joint distribution:

X\Y	0	1
0	$1-\epsilon$	0.0
1	0.0	ϵ

Interpretation: the degree of association between X,Y is very very high. In fact, one can perfectly fit a straight line, so $\rho(X,Y) = 1$. However, $I(X;Y) = H(Y) = H(1-\epsilon, \epsilon)$. As $\epsilon \to 0$, mutual information gets arbitrarily close to zero.

3 Linear Learning in One Dimension (Simple Linear Regression)

Our goal is to learn a (not necessarily deterministic) mapping $f : X \to Y$. Because the mapping is typically non-deterministic, we will view (X,Y) as jointly distributed according to some distribution $p(x,y)$. Since X (the input) will be given to us, we are interested in learning $p(y|x)$ rather than $p(x,y)$. To simplify, we will focus on learning, given any x, the *expected value* of Y, namely, $\mathrm{E}[Y \mid X = x]$.

To simplify further, we will assume a *linear relationship* between X and $\mathrm{E}[Y]$:

$$\mathrm{E}[Y \mid X] = \alpha + \beta X$$

Or, equivalently:

$$Y = \alpha + \beta X + \epsilon$$

where ϵ is some zero-mean distribution.

This is called a linear model. α, β are the parameters of the model. β is the slope, α is the intercept, or offset.

Given a set of data $\{X_i, Y_i\}_{i=1}^n$, how should we estimate the parameters α, β?

For any given values of α, β, we can plot the line on top of the datapoints, and consider the 'errors', or *residuals*:

$$\epsilon_i = y_i - (\alpha + \beta x_i)$$

I say 'error' in quotes because these are not necessarily errors! They are just the difference between the observed value of Y_i and the expected value of $Y|X = x_i$.

One possible criterion for fitting the parameters: find the parameter values that minimize the sum of squared residuals. We will see later that this corresponds to assuming that the 'error' is Guassian (Normally distributed).

This choice, called the "ordinary least squares (OLS) solution", can be written as:

$$(\alpha, \beta)_{OLS} = \arg\min_{\alpha,\beta} \sum_i (\epsilon_i)^2 = \arg\min_{\alpha,\beta} \sum_i (Y_i - (\alpha + \beta X_i))^2$$

This has a closed-form solution:

$$\beta_{OLS} = (\overline{XY} - (\overline{X})(\overline{Y}))/(\overline{X^2} - (\overline{X})^2) = \text{Cov}(X, Y)/\text{Var}(X)$$

$$\alpha_{OLS} = \overline{Y} - \beta\overline{X}$$

where $\overline{X} = \frac{1}{n}(\sum_i X_i)$, and Cov, Var are considered over the *empirical distribution*, namely the dataset.

4 Linear Learning in Multiple Dimensions

Again we want to learn a (not necessarily deterministic) mapping, but this time of the form

$$f : X_1, X_2 \ldots X_p \to Y$$

Note: X_j is now the j^{th} **random variable**, *not* the j^{th} data point!

We will again assume a *linear relationship* between $\text{E}[Y]$ and all the X variables:

$$\text{E}[Y \mid X_1, \ldots X_p] = \alpha + \sum_{j=1}^{p} \beta_j X_j$$

Or, equivalently:

$$Y = \alpha + \sum_{j=1}^{p} \beta_j X_j + \epsilon$$

where ϵ is some zero-mean distribution.

This is the linear model in multiple dimensions. α and β_j's are the parameters of the model. β_j is the slope along dimension j, α is the intercept, or offset.

To simplify notation, we define $\beta_0 = \alpha$, and set X_0 to be a 'dummy' RV with a constant value of 1. The model can then be written more succinctly as:

$$Y = \sum_{j=0}^{p} \beta_j X_j + \epsilon$$

We now have $p + 1$ parameters instead of only 2. Given a set of data, how should we estimate these parameters?

Each data point now consists of $p+1$ real valued attributes, corresponding to $X_1, X_2 \dots X_p$ and Y. Our dataset can be represented in a matrix form:

$$\mathbf{x} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

We use i (row) as index over training instances (tokens), and j (column) as index over the different attributes (a.k.a. covariates, observations, features, predictors, regressors, and independent variables). Y is called the *response variable* (a.k.a. regressand, dependent variable, and label).

For any given values of the β's, we will again consider the residuals:

$$\epsilon_i = y_i - (\sum_{j=0}^{p} \beta_j x_{i,j})$$

We can use the same criterion we used in the one-dimensional case: find the parameter values that minimize the sum of squared residuals. This choice can be written as:

$$\overline{\beta}_{OLS} = \arg\min_{\overline{\beta}} \sum_i (\epsilon_i)^2 = \arg\min_{\overline{\beta}} \sum_i (\mathbf{y}_i - (\sum_{j=0}^{p} \beta_j \mathbf{x}_{i,j}))^2 = \arg\min_{\overline{\beta}} ||\mathbf{y} - \mathbf{x}\beta||_2$$

And here too, just like in the single variable case, there is a closed-form solution:

$$\overline{\beta}_{OLS} = (\mathbf{x}^T\mathbf{x})^{-1}(\mathbf{x}^T\mathbf{y})$$

where $\mathbf{x}$ is the input matrix, and $\mathbf{y}$ is the corresponding output vector.

What is the computational complexity of calculating this solution?

- Multiplying matrices of dimensions (q, r) and (r, s) takes $O(qrs)$ operations.
- Multiplying square (n, n) matrix takes $O(n^3)$ by the naive algorithm, but only $O(n^{2.373})$ using the currenlty fastest known algorithm (Virginia Vassilevska Williams)
- Inverting a square matrix takes $O(n^{2.373})$ by the fastest known algorithm.

And so:

- $\mathbf{x}$ is a (n, p) matrix. So $\mathbf{x}^T$ has dimensions (p, n).
- $\mathbf{x}^T\mathbf{x}$ is (p, p) and takes $O(np^2)$ to calculate.
- $(\mathbf{x}^T\mathbf{x})^{-1}$ is of dimension (p, p) and takes an additional $O(p^{2.373})$ to calculate.
- $\mathbf{y}$ has dimensions $(n, 1)$. $\mathbf{x}^T\mathbf{y}$ is a $(p, 1)$ vector and takes $O(pn)$ to calculate.
- β is $(p, p) \times (p, 1) = (p, 1)$ vector and takes an additional $O(p^2)$ to calculate.

- Alternatively, $\mathbf{x}^* = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T$ is $(p,p) \times (p,n) = (p,n)$ and takes an additional $O(np^2)$ to calculate.

So overall asymptotic complexity is $O(np^2 + p^{2.373})$. Namely, for a fixed number of covariates p, the algorithm is linear in the number of datapoints. This is just about as good as it ever gets!

Is $\mathbf{x}^T\mathbf{x}$ always invertible (non-singular)? Yes, as long as no feature is a linear combination of some other features, i.e. $\mathbf{x}$ is full column rank. When $n >> p$, this is usually the case. In that case, $\mathbf{x}^* = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T$ is the left-inverse of $\mathbf{x}$, in that $\mathbf{x}^*\mathbf{x} = I$. This is a special case of a pseudo-inverse of a matrix. (If some feature is a linear combination of other features, there is no unique solution, and we say that the corresponding βs are non-identifiable)

Even if $\mathbf{x}^T\mathbf{x}$ is invertible, we may prefer to calculate $\mathbf{x}^*$ directly, because inverting a matrix can be numerically unstable.

Let's stop and consider:

1. What is the hard bias of the linear model?
2. What is the soft bias of the OLS solution?
3. What is the computational complexity of learning under these assumptions?

4.1 What Exactly is "Linear" in a Linear Model?

item The model described above is called linear because it is **linear in its parameters**. The covariates need not be linear in the original inputs/observations. In fact one could have, e.g. $X_1 = X, X_2 = X^2, X_3 = \cos(X)$, or even combinations of multiple inputs $X, Y \ldots$. So very non-linear effects can be incorporated into the linear model!

Special case: $X_j = X^j$. This is called "Polynomial regression".

4.2 Sparse Estimation, and Regularization

We focused on the case where $n >> p$. Sometimes we are in the opposite situation ($p >> n$), which we call *sparse estimation*. In that case, the OLS problem is non-identifiable: there are many solutions that match the training data equally well. So we may want to introduce a soft bias among them. Some reasonable choices:

- Minimize the L2 norm of the βs (sum of squared βs). This is called "Ridge regression".
- Minimize the L1 norm of the βs (sum of absolute values of the βs). This is called "Lasso Regression".
- Minimize the L0 norm of the βs (Number of non-zero βs, namely number of contributing covariates). This is called "Subset Regression". It makes a lot of scientific sense (i.e. compact explanations in terms of few covariates), but unfortunately it is computationally intractable to optimize for large problems.

These are all cases of *regularization*: a generalization of the Occam's Razor principle, where we try to balance the fit to the data with some prior preference (soft bias) over the set of solutions.