More Probability, Decision Theory, and the Bias-Variance Tradeoff

Rob McCulloch

1. Continuous Random Variables
2. Expection, Mean, Variance, Covariance
3. Statistical Decision Theory
4. The Bias Variance Tradeoff
5. The Behaviour of a Mean

1. Continuous Random Variables

Sometimes it is inconvenient to list out all the possible values a random variable can take on.

For example, we don't want to list all the possible times a patient could live for.

In this case we let our random variables take on on value in R, or any value in a subset of R.

For example we might think of the time our patient live to be any value in the subset of R given by $\{x; x > 0\}$.

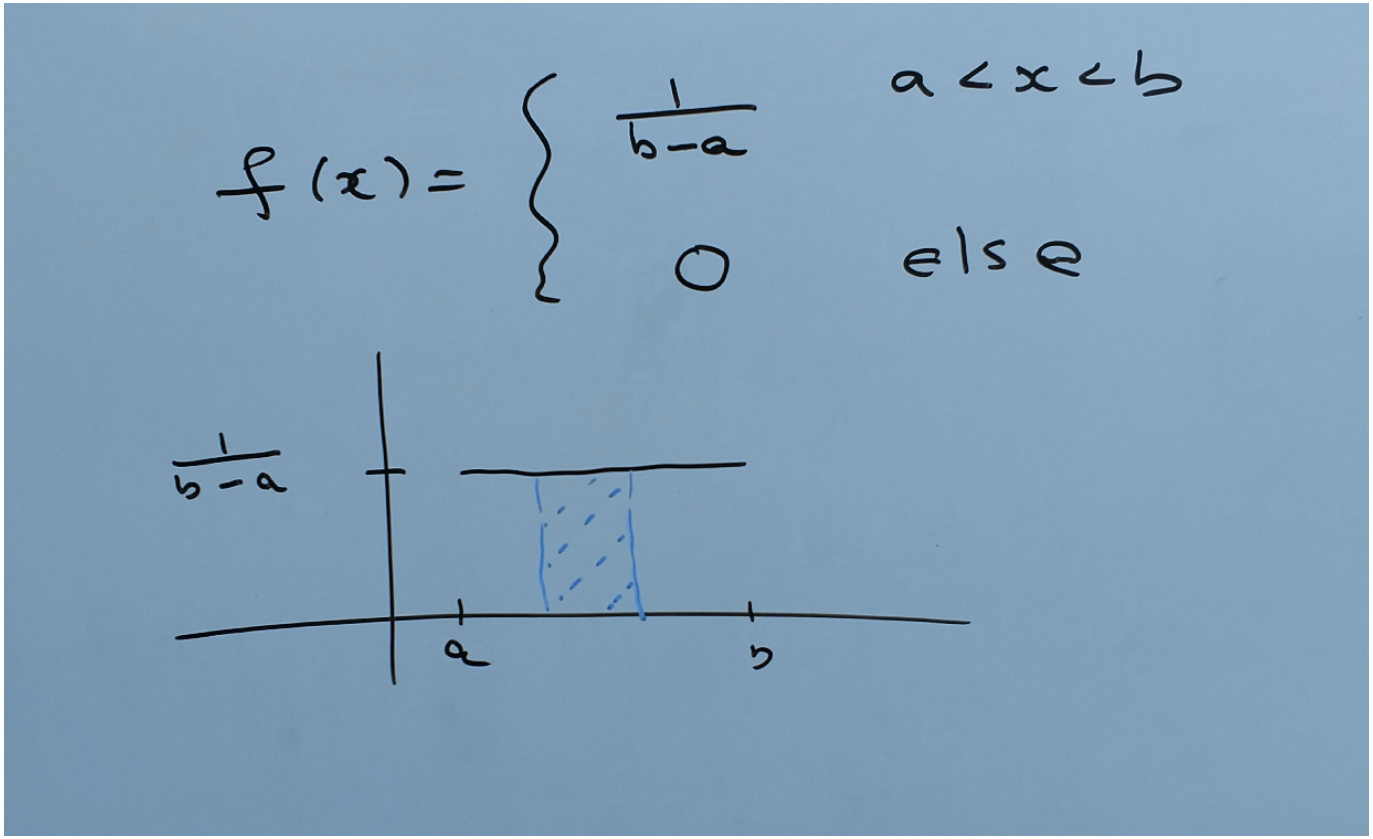
In this case our random variable (or vector) is a *continuous random variable*.

For continuous random variable we don't talk about the probability of a particular value, we can only talk about about the probability of a set.

We use the *probability density function* (*pdf*) f_x to specify the probability of a set A by

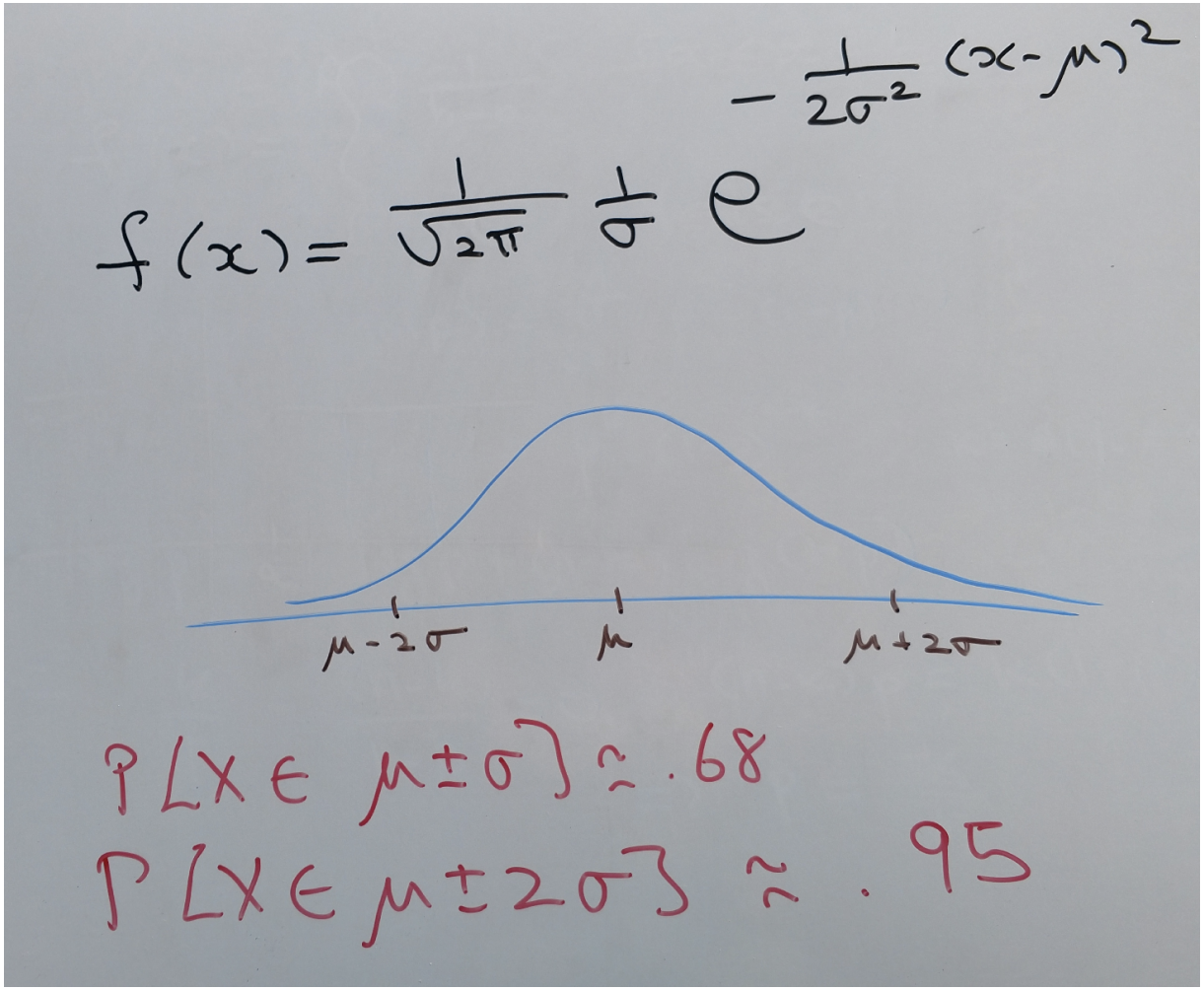
$$p(X \in A) = \int_A f_x(x)\, dx$$

Example, the Uniform



We write $X \sim U(a, b)$.

Example, the Normal



We write $X \sim N(\mu, \sigma^2)$.

Basic Properties

The basic properities we had in the discrete case extend to the continous case:

$$f(y_1, y_2, y_3, \ldots, y_n) = f(y_1)\, f(y_2 \mid y_1)\, f(y_3 \mid y_1, y_2)\, f(y_n \mid y_1, y_2, \ldots, y_{n-1})$$

If the Y_i are independent then

$$f(y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} f(y_i)$$

Margining out:

$$f(y_1) = \int f(y_1, y_2)\, dy_2$$

Conditional:

$$f(y_1 \mid y_2) = \frac{f(y_1, y_2)}{f(y_2)}$$

Bayes theorem:

$$\begin{aligned} f(y_2 \mid y_1) &\propto f(y_2)\, f(y_1 \mid y_2) \\ &= f(y_1, y_2) \end{aligned}$$

2. Expection, Mean, Variance, Covariance

Let Y be a random variable (or vector).

Sometimes we want to summarize the possible values of some function of Y.

We use a probability weighted average:

Discrete:

$$E(h(Y)) = \sum h(y)p(y)$$

Continuous:

$$E(h(Y)) = \int h(y)\, f(y)\, dy$$

The key examples are the mean and variance of a univariate random variable.

The Mean:

$h(y) = y$

$$\begin{aligned} E(Y) &= \sum y\, p(y) \quad \text{(discrete)} \\ &= \int y\, f(y)\, dy \quad \text{(continuous)} \end{aligned}$$

We often write μ or μ_y for $E(Y)$.

The Variance: $h(y) = (y - \mu)^2$.

$$\begin{aligned} Var(Y) &= \sum (y-\mu)^2 \, p(y) \quad \text{(discrete)} \\ &= \int (y-\mu)^2 \, f(y) \, dy \quad \text{(continuous)} \end{aligned}$$

We often write σ^2 or σ_y^2 for $Var(Y)$.

The Standard Deviation:

$$\sigma = \sqrt{(\sigma^2)}$$

is the *standard deviation*.

Note that σ has the same units as Y.

The variance and standard deviation summarize how close a random variable tends to be to its mean.

Example, the Bernoulli:

$X \sim$ Bernoulli(p) means:

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

$E(X) = (1 - p) \times 0 + p \times 1 = p.$

$$\begin{aligned} Var(X) &= \sum_{i=1}^{n} P(x_i) \times [x_i - E(X)]^2 \\ &= (1 - p) \times (0 - p)^2 + p \times (1 - p)^2 \\ &= p(1 - p) \times [p + (1 - p)] \\ Var(X) &= p(1 - p) \end{aligned}$$

Example, the Normal:

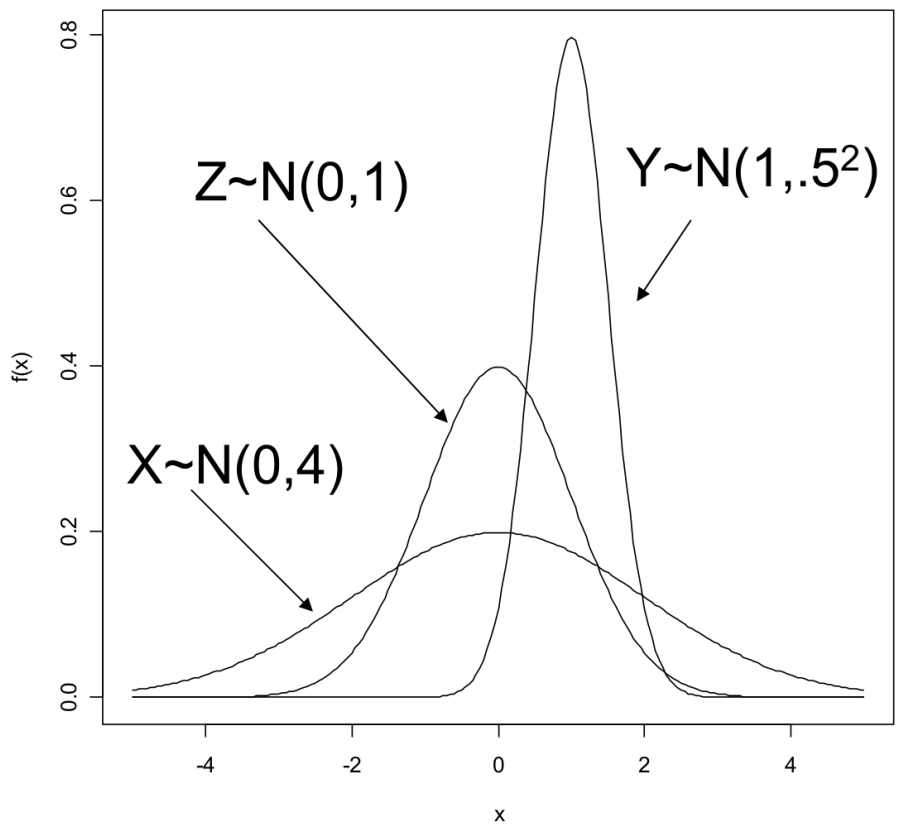
You can show that for $X \sim N(\mu, \sigma^2)$,

$$E(X) = \mu, \quad Var(X) = \sigma^2, \quad \sigma_X = \sigma.$$

The normal family has two parameters

μ: where the curve is centered

σ: how spread out the curve is



A small σ means the distribution is "tight" around μ !!

Expectation of Linear Functions:

If $h(X, Y) = a + b\,X + c\,Y$ then,
$E(h(X, Y)) = a + b\,E(X) + c\,E(Y)$.

$$\begin{aligned} E(h(X,Y)) &= \sum_{(x,y)} p(x,y)(a + bx + cy) \\ &= a\sum p(x,y) + b\sum x\,p(x,y) + c\sum y\,p(x,y) \\ &= a + b\sum x\,p(x) + c\sum y\,p(y) \\ &= a + bE(X) + cE(Y) \end{aligned}$$

More generally,

$$E(a + \sum b_i\,Y_i) = a + \sum b_i E(Y_i)$$

Covariance and Correlation:

The covariance and correlation are used to measure how much one random variable looks like a linear function of another.

Let $E(Y_1) = \mu_1$ and $E(Y_2) = \mu_2$.

$h(y_1, y_2) = (y_1 - \mu_1)(y_2 - \mu_2)$.

$$Cov(Y_1, Y_2) = E((Y_1 - \mu_1)(Y_2 - \mu_2)).$$

We might write $\sigma_{X,Y}$ for $Cov(X, Y)$, or σ_{12} for $Cov(Y_1, Y_2)$.

The Correlation

Let σ_i be the standard deviation of Y_i.

$$Cor(Y_1, Y_2) = \frac{\sigma_{12}}{(\sigma_1 \, \sigma_2)}.$$

The covariance divided by the product of the the standard deviations.

We might write ρ_{XY} for $Cor(X, Y)$ for ρ_{12} for $Cor(Y_1, Y_2)$.

Key Property of Correlation:

$$-1 \leq \rho_{X,Y} \leq 1$$

The correlation is always between 1 and -1.

The closer the correlation is to 1, the more $Y \approx a + bX$ with $b > 0$.

The closer the correlation is to -1, the more $Y \approx a + bX$ with $b < 0$.

Independence, Expectation, and Correlation:

Suppose X and Y are independent random variables.

Then $E(XY) = E(X)\,E(Y)$.

$$
\begin{aligned}
E(XY) &= \int\int xy\, f(x,y)\, dxdy \\
&= \int \left(\int xy\, f(x) f(y)\, dy \right) dx \\
&= \int x\, f(x) \left(\int y\, f(y)\, dy \right) dx \\
&= \left(\int x\, f(x) dx \right) \left(\int y\, f(y) dy \right) \\
&= E(X)\, E(Y)
\end{aligned}
$$

In the discrete case with $X \in \{x_1, x_2\}$ and $Y \in \{y_1, y_2\}$ we have

$$E(X)E(Y) = (p(x_1)x_1 + p(x_2)x_2)(p(y_1)y_1 + p(y_2)y_2)$$

$$= (p(x_1)x_1p(y_1)y_1 + p(x_1)x_1p(y_2)y_2 +$$

$$\quad (p(x_2)x_2p(y_1)y_1 + p(x_2)x_2p(y_2)y_2 +$$

$$= \sum x_i x_j p(x_i)p(y_j)$$

$$= \sum x_i x_j p(x_i, y_j)$$

$$= E(XY)$$

Suppose X and Y are independent.

Then,

$$\begin{aligned} \sigma_{XY} &= E((X-\mu_X)(Y-\mu_y)) \\ &= E(X-\mu)E(Y-\mu) \\ &= 0 \times 0 = 0 \end{aligned}$$

Independent $\Rightarrow \rho_{XY} = 0$, but not the other way around.

Variance of a linear function

Suppose $Z = a + b\,X + c\,Y$.

$Z - E(Z) =$ b (X-E(X)) + c (Y-E(Y)).

$(Z - E(Z))^2 =$
$b^2(X - E(X))^2 + c^2(Y - E(Y))^2$
$\quad +bc\,(X - E(X))(Y - E(Y)) + bc\,(X - E(X))(Y - E(Y))$.

$Var(Z) = b^2\,Var(X) + c^2\,Var(Y) + 2bc\,Cov(X, Y)$.

More generally, if $Y = a + \sum b_i X_i$ then

$$Var(Y) = \sum b_i^2 Var(X_i) + \sum_{i \neq j} b_i b_j Cov(X_i, X_j).$$

$$= \sum b_i^2 Var(X_i) + 2 \sum_{i<j} b_i b_j Cov(X_i, X_j).$$

Variance (Covariance) Matrix of a Random Vector

Suppose $X = (X_1, X_2, \ldots, X_p)$.

Let $\sigma_{ii} = Var(X_i)$ and $\sigma_{i,j} = Cov(X_i, X_j)$.

Then the matrix with (i, j) element equal to σ_{ij} is the variance matrix, or variance-covariance matrix of X.

We often use Σ to denote this matrix.

$$\Sigma = [\sigma_{ij}].$$

Mean and Variance of a Linear Function with Matrix Notation

Let $X = (X_1, X_2, \ldots, X_p)'$ and $b = (b_1, b_2, \ldots, b_p)' \in R^p$.

Let $E(X) = (E(X_1), E(X_2), \ldots, E(X_p))' = \mu'$
We have,

$$E(a + b'X) = a + b'\mu.$$

And,

$$Var(a + b'X) = b'\,\Sigma\, b$$

Iterated Expectations:

Let $f(X, Y)$ be a function of the random variables X and Y.

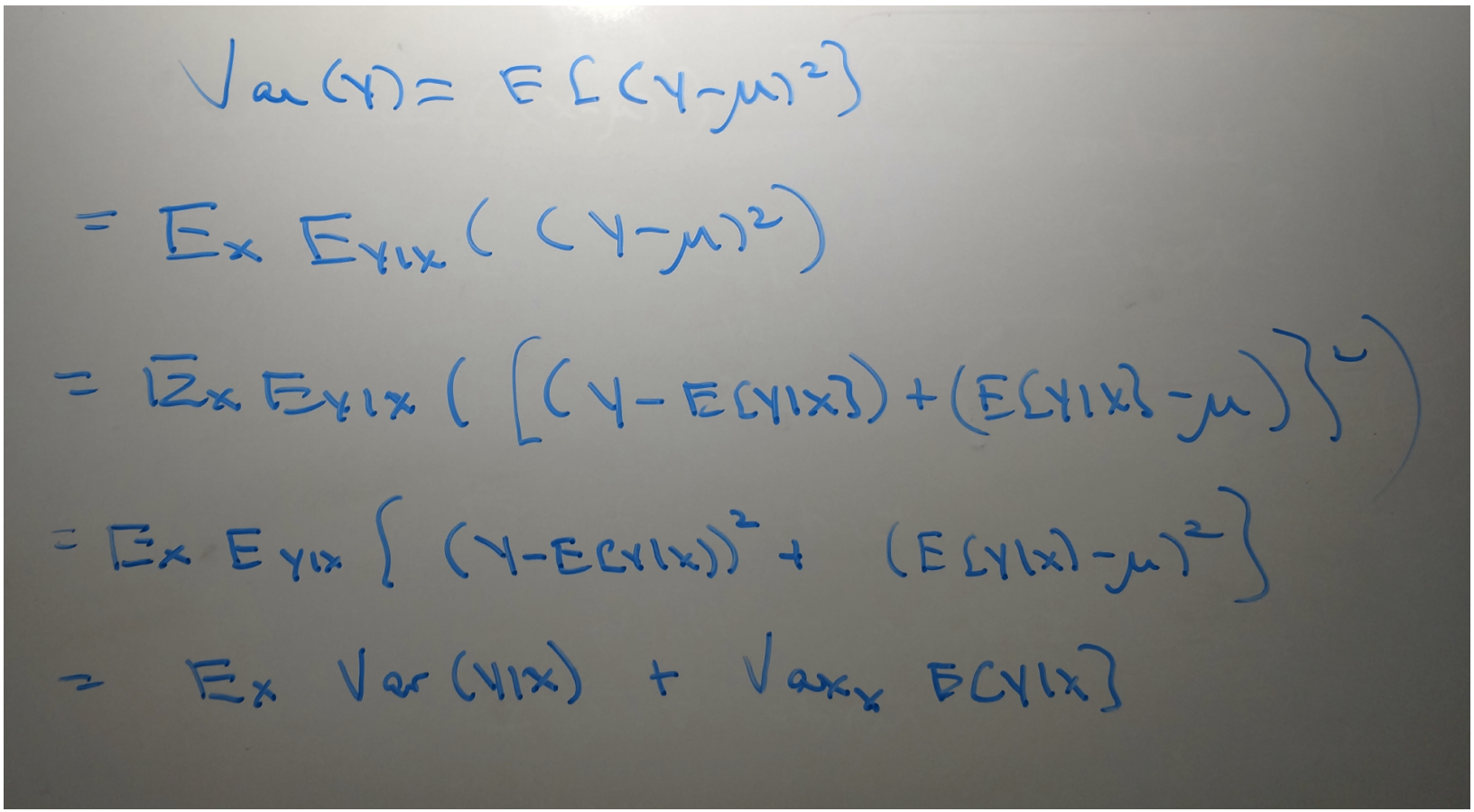
Let $p(x, y)$ be the joint density.

$$\begin{aligned}
E(f(X,Y)) &= \int f(x,y)p(x,y)\,dxdy \\
&= \int f(x,y)p(y \mid x)p(x)\,dxdy \\
&= \int p(x)[\int f(x,y)p(y \mid x)\,dy]\,dx \\
&= \int E(f(x,Y) \mid x)\,p(x)dx \\
&= E_X(E_{Y|X}(f(X,Y)))
\end{aligned}$$

For example, $E(Y) = E_X(E(Y \mid X))$.

Example:

This one comes up from time to time:



3. Statistical Decision Theory

See section 2.4, Elements of Statistical Learning, "Statistical Decision Theory".

Suppose we want to guess the random variable Y.
Let m be a possible guess and $\mu = E(Y)$.

Let's choose the guess that minimizes the expected squared error:

$$
\begin{aligned}
& E((Y-m)^2) \\
= \; & E((Y-\mu)+(\mu-m))^2) \\
= \; & E\{(Y-\mu)^2 + 2(Y-\mu)(\mu-m) + (\mu-m)^2\} \\
= \; & \mathrm{Var}(Y) + 0 + (\mu-m)^2 \\
= \; & \mathrm{Var}(Y) + (\mu-m)^2
\end{aligned}
$$

So, the best choice is $m^* = \mu$.

Now suppose we have (X, Y) where X is a random vector and Y is a random variable.

Let $\mu(x) = E(Y \mid X = x)$.

We want to minimize the expected squared error if we predict Y using $f(X)$.

$$E((Y - f(x))^2)$$

$$= E_X \, E_{Y|X}\left((Y - f(x))^2 \mid X = x\right)$$

$$= E_X\left(\mathrm{Var}(Y|x) + (\mu(x) - f(x))^2\right)$$

Clearly, the optimal *function* is $f^*(x) = \mu(x)$.

Minimizing Expected Loss

In general, we specify a loss function $L(y, a(x))$.

Given the information in x we can choose an *action*, and we incur a *loss* which depends on our action and the outcome y.

We then minimize the the expected loss, where the expectation is taken over the joint distribution of (X, Y):

$$\underset{a}{\text{minimize}}\ E(L(Y, a(X))$$

We just did $L(y, a(x)) = (y - a(x))^2$.

Note:

In principle, we should pick losses that are meaningful given the actual applied setting.

As a practical matter, we often use choose generic loss functions in order to devise general methods.

We just used the generic loss $L(y, a(x)) = (y - a(x))^2$ which is by far the most commonly used loss for a numeric outcome.

For example if $L(y, a(x)) = |y - a(x)|$ then the optimal action is the median of the conditional distribution $Y \mid X = x$ rather than the mean.

4. The Bias Variance Tradeoff

We have studied the bias-variance tradeoff informally.

Following ISLR, section 2.2.2 and ESL section 7.2, we can formalize this a bit.

We will assume the model

$$Y = f(X) + \epsilon$$

where ϵ is independent of X.

Our setup is that we have training data (x_i, y_i) generated under our model so that $Y_i = f(x_i) + \epsilon_i$.

We are predicting at a specific x_0 and

$$Y_0 = f(x_0) + \epsilon$$

where ϵ is independent of the training data.

We want to consider the variation due the training data (X_i, Y_i) and the variation due to Y_0 (which is due to ϵ).

Given the training date we assume we have an algorithm for generating $\hat{f}$ and estimate of f.

Given the particular choice x_0, we can then think of $\hat{f}(x_0)$ as a random variable where the variation is driven by the random training data.

We consider

$$E(Y_0 - \hat{f}(x_0))^2.$$

The expection is over the random variable Y_0 (driven by ϵ) and the random variable $\hat{f}(x_0)$ (driven by the training data). These two random variable are independent.

Example:

If our statistical learner is multiple regression then our model is

$$Y = x'\beta + \epsilon$$

and

$$\hat{f}(x_0) = x_0'\hat{\beta},$$

where $\hat{\beta}$ is the usual regression estimator of β.

Example:

If we use KNN for a fixed k, then we do get a $\hat{f}(x_0)$ that depends on the training data and this is what we simulated in our previous notes.

However KNN does not fit into the $Y = f(X) + \epsilon$ setup and we don't have a simple parameter related restricting the choice of f.

Note:

If E(Y) =0 then

$$\begin{aligned} E((a+Y)^2) &= E(a^2+2aY+Y^2) \\ &= a^2+2aE(Y)+E(Y^2) \\ &= a^2+Var(Y) \end{aligned}$$

Note:

If X and Y are uncorrelated (in particular, if they are independent) then

$$Var(X+Y) = Var(X)+Var(Y)$$

$$E(Y_0 - \hat{f}(x_0))^2 = E[((f(x_0) - \hat{f}(x_0)) + \epsilon)^2]$$
$$= E[((f(x_0) - E(\hat{f}(x_0)) - (\hat{f}(x_0) - E(\hat{f}(x_0)) + \epsilon)^2]$$
$$= (f(x_0) - E(\hat{f}(x_0))^2 + Var(\hat{f}(x_0)) + Var(\epsilon)$$

$=$ (the bias squared) $+$ (the variance) $+$ (irreducible error)

The "irreducible error" is the variation in the part of Y_0 we cannot learn from X.

Even if we knew f, we would still have this.

If we make $\hat{f}$ more complicated we expect the bias to go down and the variance to go up !!!!!!!

5. The Behaviour of a Mean

We do a lot of averaging !!!

Suppose W_i are iid with $E(W_i) = \mu$ and $Var(W_i) = \sigma^2$.

$$\bar{W} = \frac{1}{n} \sum W_i.$$

Then,

$$\begin{aligned} E(\bar{W}) &= \frac{1}{n} \sum E(W_i) \\ &= \frac{1}{n} n\mu \\ &= \mu \end{aligned}$$

$$
\begin{aligned}
Var(\bar{W}) &= \frac{1}{n^2}\sum Var(W_i) \\
&= \frac{1}{n^2} n\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

So, as n gets big, the distribution of $\bar{W}$ tightens up around $E(W)$!!

In fact, by the central limit theorem,

$$\bar{W} \approx N(\mu, \sigma^2/n)$$

so that,

$$\bar{W} \approx \mu \pm 2\frac{\sigma}{\sqrt{n}}$$

with 95% probability.

In theory, we have $W = L(Y, a(X))$ and we want an action function $a(X)$ which makes our expected loss small:

$$E(W) = E(L(Y, a(X)).$$

In practice, we estimate this with

$$\bar{W} = \frac{1}{n} \sum L(Y_i, a(X_i))$$

where we average over the test data.

In practice we have to use data to come up with our action, see we also have the training data:

$$T = (X^t, Y^t)$$

Then, we "learn" our action using the training data:

$$\hat{a}(x, \gamma) = a(x, T, \gamma)$$

where γ is a tuning parameter (e.g k in KNN).

We then have a loss which depends on our action and the future Y.

We get to see X to help us guess Y:

$$L(Y, \hat{a}(X, \gamma))$$

we want an action scheme that minimizes

$$E_{Y,X,T} L(Y, \hat{a}(X, \gamma))$$

Cross validation is supposed to estimate this expectation!!!

So, the size of the test data help us estimate the expecation over (X, Y).

The observations within the test data are independent.

The number of folds is the number of draws we have of the training data.

Are they indepedent?

Why is leave one out at at time (loocv) such a bad idea?