



THE  
POWER  
TO KNOW.

**SAS<sup>®</sup> Enterprise Miner<sup>™</sup> and  
SAS<sup>®</sup> Text Miner Procedures  
Reference for SAS<sup>®</sup> 9.1.3  
The DECIDE Procedure  
(Book Excerpt)**

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., SAS® *Enterprise Miner™* and SAS® *Text Miner Procedures: Reference for SAS® 9.1.3*, Cary, NC: SAS Institute Inc.

**SAS® Enterprise Miner™ and SAS® Text Miner Procedures: Reference for SAS 9.1.3**

Copyright © 2008 by SAS Institute Inc., Cary, NC, USA.

All rights reserved. Produced in the United States of America.

**For a Web download or e-book:** Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

**U.S. Government Restricted Rights Notice.** Use, duplication, or disclosure of this software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227-19 Commercial Computer Software-Restricted Rights (June 1987).

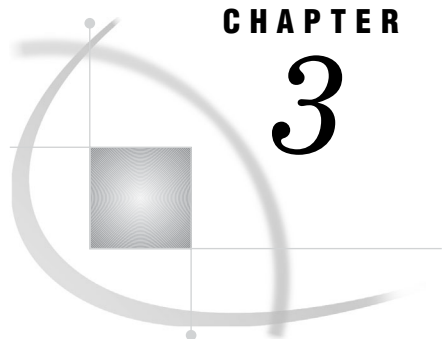
SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st electronic book, October 2008

SAS® Publishing provides a complete selection of books and electronic products to help customers use SAS software to its fullest potential. For more information about our e-books, e-learning products, CDs, and hard-copy books, visit the SAS Publishing Web site at [support.sas.com/publishing](http://support.sas.com/publishing) or call 1-800-727-3228.

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.



## CHAPTER

## 3

## The DECIDE Procedure

---

<i>Overview: DECIDE Procedure</i>	39
<i>Syntax: DECIDE Procedure</i>	40
<i>PROC DECIDE Statement</i>	40
<i>CODE Statement</i>	41
<i>DECISION Statement</i>	42
<i>FREQ Statement</i>	44
<i>POSTERIOR Statement</i>	45
<i>PREDICTED Statement</i>	45
<i>TARGET Statement</i>	46
<i>Details: DECIDE Procedure</i>	46
<i>Example: DECIDE Procedure</i>	46
<i>Example 1: Using the DECIDE Procedure Following the DISCRIM Procedure</i>	46
<i>References</i>	50

---

### Overview: DECIDE Procedure

The DECIDE procedure produces optimal decisions based on a user-supplied decision matrix, prior probabilities, and output from a modeling procedure. The output from the modeling procedure can be either posterior probabilities for the classes of a categorical target variable or predicted values of an interval target variable. The DECIDE procedure can also adjust posterior probabilities for changes in the prior probabilities. Background and formulas for decision processing are given in the documentation on “Predictive Modeling.”

The decision matrix contains columns (decision variables) corresponding to each decision and rows (observations) corresponding to target values. The values of the decision variables represent target-specific consequences, which might be profit, loss, or revenue. These consequences are the same for all cases being scored.

For a categorical target variable, there should be one row for each category. The value in the decision matrix that is located at a given row and column specifies the consequence of making the decision corresponding to column when the target value corresponds to the row.

For an interval target variable, each row defines a knot in a piecewise linear spline function. The consequence of making a decision is computed by interpolating in the corresponding column of the decision matrix. If the predicted target value is outside the range of knots in the decision matrix, the consequence of a decision is computed by linear extrapolation.

For each decision, there might also be either a cost variable or a numeric constant. The values of these variables represent case-specific consequences, which are always

costs. These consequences do not depend on the target values of the cases being scored. Costs are used for computing return on investment as  $(\text{revenue}-\text{cost})/\text{cost}$ .

Cost variables might be specified only if the decision data set contains revenue, not profit or loss. Therefore, if revenues and costs are specified, profits are computed as revenue minus cost. If revenues are specified without costs, the costs are assumed to be 0. The interpretation of consequences as profits, losses, revenues, and costs is needed only to compute return on investment. You can specify values in the decision data set that are target-specific consequences but that might have some practical interpretation other than profit, loss, or revenue. Likewise, you can specify values for the cost variables that are case-specific consequences but that might have some practical interpretation other than costs. If the revenue/cost interpretation is not applicable, the values computed for return on investment might not be meaningful.

The DECIDE procedure will choose the optimal decision for each observation. If the decision data set is of TYPE=PROFIT or REVENUE, the decision that produces the maximum expected or estimated profit is chosen. If the decision data set is of TYPE=LOSS the decision that produces the minimum expected or estimated loss is chosen.

If the actual value of the target variable is known, the DECIDE procedure will calculate:

- The consequence of the chosen decision for the actual target value for each case.
- The best possible consequence for each case.
- Summary statistics giving the total and average profit or loss.

Some modeling procedures assume that the prior probabilities for categorical variable level membership are either all equal or proportional to the relative frequency of the corresponding response level in the data set. PROC DECIDE allows you to specify other prior probabilities. Thus, you can conduct a sensitivity analysis without rerunning the modeling procedure.

---

## Syntax: DECIDE Procedure

```
PROC DECIDE <option(s)>;
  CODE <option(s)>;
  DECISION DECDATA=<libref.> SAS-data-set <option(s)>;
  FREQ variable;
  POSTERiors variable(s);
  PREDICTED variable;
  TARGET variable;
```

---

## PROC DECIDE Statement

Invokes the DECIDE procedure.

**Discussion:** The PROC DECIDE statement runs the DECIDE procedure and identifies the input and output data sets. You also need the following statements:

- A DECISION statement
  - Either a POSTERiors or a PREDICTED statement
  - A TARGET statement
-

**PROC DECIDE** *<option(s)>*;

## Options

**DATA=***<libref.>SAS-data-set*

Specifies the input data set that contains the output from a modeling procedure.

**Default:** `_LAST_`

**OUT=***<libref.>SAS-data-set*

Specifies the output data set that contains the following variables:

- the variables from the input data set
- the chosen decision (prefix `D_`)
- the expected consequence of the chosen decision (prefix `EL_` or `EP_`)

If the target value is in the input data set, the output data set also contains the following variables:

- the consequence of the chosen decision computed from the target value (prefix `CL_` or `CP_`)
- the consequence of the best possible decision knowing the target value (prefix `BL_` or `BP_`)

If `PRIORVAR=` and `OLDPRIORVAR=` variables are specified, then the output data will contain the recalculated posteriors.

*Note:* If you want to create a permanent SAS data set, you must specify a two-level name. For more information on this topic, see “SAS Files” and “DATA Step Concepts “ in *SAS Language Reference: Concepts*.  $\Delta$

**Default:** If the `OUT=` option is omitted, PROC DECIDE creates an output data set and names it according to the `DATA $n$`  convention, just as if you had omitted a data set name in a DATA statement.

**OUTFIT=***<libref.> SAS-data-set*

Specifies the output data set that contains statistics including the total and average profit or loss. The `OUTFIT=` option might not be specified with `ROLE=SCORE`.

**Default:** None

**ROLE=**`TRAIN | VALID | VALIDATION | TEST | SCORE`

Specifies whether the `DATA=` data set is a training set, validation set, test set, or scoring set. The `ROLE=` option affects the names of the variables in the `OUTFIT=` data set.

**Default:** `TEST`

---

## CODE Statement

**Generates SAS DATA step code that can be used to score data sets.**

**Tip:** If neither `FILE=` nor `METABASE=` are specified, then the SAS code is written to the SAS log. You can specify both `FILE=` and `METABASE=` to write code to both locations. The `TARGET` variable must appear in the `DATA=` data set as well as the `DECDATA=` data set.

---

**CODE** *<code-option(s)>*;

## Code Options

### **FILE=***'filename'*

Specifies a path for writing the code to an external file. For example:

FILE="c:\mydir\scorecode.sas".

**Default:** None

### **FORMAT=***format*

Specifies the numeric format to be used when printing numeric constants. For example, FORMAT=BEST20.

**Default:** FORMAT=BEST12

### **GROUP=** *group-name*

Specifies the group identifier (up to four characters) for group processing.

**Default:** GROUP=\_

### **METABASE=***screenspec*

Specifies a catalog entry to which the code is written. For example, METABASE=mylib.mycat.myentry.

**Default:** None

### **RESIDUAL**

Specifies that variables that depend on the target variable, such as the BL\_, BP\_, CL\_, and CP\_ variables, are to be computed in the code.

---

## DECISION Statement

Specifies information used for decision processing in the **DECIDE**, **DMREG**, **NEURAL**, and **SPLIT** procedures. *This documentation applies to all four procedures.*

**Tip:** The **DECISION** statement is required in the **DECIDE** procedure, but not in the **DMREG**, **NEURAL**, or **SPLIT** procedures.

---

**DECISION DECADATA=***<libref.>SAS-data-set <option(s)>*;

### **DECADATA=** *<libref.>SAS-data-set*

Specifies the input data set that contains the decision matrix and/or prior probabilities. The DECADATA= data set must also contain the target variable.

The DECADATA= data set might contain decision variables specified by means of the **DECVAR=** option, or prior probability variable(s) specified by means of the **PRIORVAR=** option and/or the **OLDPRIORVAR=** option, or both.

The target variable is specified by means of the **TARGET** statement in the **DECIDE**, **NEURAL**, and **SPLIT** procedures or by using the **MODEL** statement in the **DMREG** procedure.

For a categorical target variable, there should be one row for each class. The value in the decision matrix located at a given row and column specifies the consequence of

making the decision corresponding to column when the target class corresponds to the row. If any class appears twice or more in the DECADATA= data set, an error message is printed and the procedure terminates. For the DMREG, NEURAL, and SPLIT procedures, all class values in the training set must also appear in the DECADATA= data set, but it is allowed to have class values in the DECADATA= data set that are not in the training set. For the DECIDE procedure, any class value in the DATA= data set that is not found in the DECADATA= data set is treated in the same way as a missing class value; it is allowed to have class values in the DECADATA= data set that are not in the DATA= data set, but note that the classes in the DECADATA= data set must correspond exactly with the variables in the POSTERiors statement.

For an interval target variable, each row defines a knot in a piecewise linear spline function. The consequence of making a decision is computed by interpolating in the corresponding column of the decision matrix. If the predicted target value is outside the range of knots in the decision matrix, the consequence of a decision is computed by linear extrapolation. If the target values are in nondecreasing or nonincreasing order, any interior target value is allowed to appear twice in the data set so you can specify discontinuities. The end points (that is, the minimum and maximum target values in the data set) might not appear more than once. No target value is allowed to appear more than twice. If the target values are not in nondecreasing or nonincreasing order, the target values are sorted by the procedure, and no target value might appear more than once.

**Tip:** The DECADATA= data set may be of TYPE=LOSS, PROFIT, or REVENUE. If unspecified, TYPE=PROFIT is assumed by default. TYPE= is a data set option that can be specified in parenthesis following the data set name when the data set is created or when the data set is used.

## Options

### **DECVARs=***decision-variable(s)*

Specifies the numeric decision variables in the DECADATA= data set that contain the target-specific consequences for each decision. The decision variables can not contain missing values. If DECVARS= is not specified, the procedure does not make any decisions or output any variables that depend on making a decision.

**Default:** None

### **COST=***cost-option(s)*

Specifies numeric constants that give the cost of a decision, or numeric variables in the DATA= data set that contain the case-specific costs, or any combination of constants and variables. There must be the same number of cost constants and variables as there are decision variables in the DECVARS= option. In the COST= option, you can not use abbreviated variable lists such as D1-D3, ABC-XYZ, or PQR.. For any case where a cost variable is missing, the results for that case are set to missing.

**Default:** All costs are assumed to be 0.

*Note:* The COST= option can be specified only when the DECADATA= data set is of TYPE=REVENUE.  $\Delta$

### **PRIORVAR=***variable*

Specifies the numeric variable in the DECADATA= data set that contains the prior probabilities to use for making decisions. Prior probabilities are also used to adjust the total and average profit or loss. Prior probabilities can not be missing or negative, and there must be at least one positive prior probability. The priors are not required to sum to 1; if they do not sum to 1, they are automatically multiplied by a

constant to do so. If PRIORVAR= is not specified, no adjustment for prior probabilities is applied to the posteriors.

**Default:** None

**OLDPRIORVAR=***variable*

Specifies the numeric variable in the DECDATA= data set that contains the prior probabilities that were used when originally fitting the model. The OLDPRIORVAR= option is used only by the DECIDE procedure. In the DMREG, NEURAL, and SPLIT procedures, the procedure automatically supplies the values of the old priors if PRIORVAR= is specified.

*Note:* If OLDPRIORVAR= is specified, PRIORVAR= must also be specified.  $\triangle$

**Default:** None

## **FREQ Statement**

**Specifies a numeric variable whose values represent the frequency of the observation.**

**Discussion:** If one variable in the input data set represents the frequency of occurrence for other values in the observation, specify the variable's name in a FREQ statement.

**Alias:** FREQUENCY

**FREQ** *variable*;

### **Required Argument**

*variable*

Specifies a single numeric variable whose value represents the frequency of the observation. If you use the FREQ statement, PROC DECIDE then treats the data set as if each observation appeared  $n$  times, where  $n$  is the value of the FREQ variable for the observation. The FREQ variable has no effect on decisions or the adjustment for prior probabilities; it only affects the summary statistics in the OUTFIT= data set. If a value of the FREQ variable is not an integer, the fractional part is *not* truncated. If a value of the FREQ variable is less than or equal to 0, the observation does not contribute to the summary statistics in the OUTFIT= data set, but all of the variables created in the OUT= data are processed the same way as if the FREQ variable were positive.

---

## POSTERIORS Statement

Specifies which variables in the DATA= data set contain the estimated posterior probabilities that correspond to the categories of the target variable.

**Discussion:** The POSTERIORS statement can be specified only with a categorical target variable. You can not use both a POSTERIORS statement and a PREDICTED statement.

---

**POSTERIORS** *variable(s)*;

### *variable(s)*

Specifies the numeric variable(s) in the DATA= data set that contain the estimated posterior probabilities corresponding to the classes (that is, the categories of the target variable). The results for a case are set to missing and the P flag is set in the \_WARN\_ variable for any case where a posterior probability is missing, negative, or greater than one, or there is a non-0 posterior corresponding to a 0 prior, or there is not at least one valid positive posterior probability.

### **CAUTION:**

The order of the variables must correspond exactly to the order of the classes in the DECDATA= data set.  $\Delta$

---

## PREDICTED Statement

Specifies which variable in the DATA= data set contains the predicted values of an interval target variable.

**Discussion:** The PREDICTED statement can be specified only with an interval target variable. You can not use both a PREDICTED statement and a POSTERIORS statement.

---

**PREDICTED** *variable*;

### *variable*

Specifies the numeric variable in the DATA= data set that contains the predicted values of an interval target variable.

---

## TARGET Statement

**Specifies which variable in the DECDATA= data set contains values for the target variable.**

**Discussion:** The DECIDE procedure will search for a target variable with the same name in the DATA= data set. If none is found, then the DECIDE procedure will assume that the actual target values are unknown. For a categorical target, the target variables in the DATA= and DECDATA= data sets need not be of the same type because the normalized formatted values are used for comparisons. For an interval target, both variables must be numeric. If scoring code is generated by a CODE statement, the code will format the target variable using the format and length from the DATA= data set.

**Tip:** The TARGET statement is required.

---

**TARGET** *variable*;

***variable***

Specifies the variable in the DECDATA= data set that contains the values for the target variable.

---

## Details: DECIDE Procedure

*Note:* Formulas for adjusting posterior probabilities and for decision processing are given in the chapter on “Predictive Modeling.”  $\Delta$

---

## Example: DECIDE Procedure

---

### Example 1: Using the DECIDE Procedure Following the DISCRIM Procedure

This example shows how to use the DECIDE procedure to adjust posterior probabilities from the DISCRIM procedure, and how to make decisions using a revenue matrix and cost constants.

In a population of men who consult urologists for prostate problems, 70% have benign enlargement of the prostate, 25% have an infection, and 5% have cancer. A sample of 100 men is taken, and two new diagnostic measures, X and Y, are made on each patient. The training set also includes the diagnosis made by reliable, conventional methods. For each patient, two treatments are available: 1) Antibiotics are effective against infection, but might have moderately bad side effects. Antibiotics have no effect on benign enlargement or cancer. 2) Surgery is effective for all diseases but has potentially severe side effects such as impotence. There is also the option of doing nothing.

*Note:* This example is purely fictional. Any resemblance to actual medical conditions or treatments is coincidental.  $\Delta$

```

data prostate;
  length dx $10;

  dx='Benign';
  mx=30; sx=10;
  my=30; sy=10;
  n=70;
  link generate;

  dx='Infection';
  mx=70; sx=20;
  my=35; sy=15;
  n=25;
  link generate;

  dx='Cancer';
  mx=50; sx=10;
  my=50; sy=15;
  n=5;
  link generate;
  stop;

generate:
  do i=1 to n;
    x=rannor(12345)*sx+mx;
    y=rannor(0) *sy+my;
    output;
  end;

run;

title2 'Diagnosis';
proc gplot data=prostate; plot y*x=dx; run;

```

Use DISCRIM to see how well inputs X and Y can classify each patient according to disease.

```

proc discrim data=prostate out=outdis short;
  class dx;
  var x y;
run;

title2 'Classification with equal priors';
proc gplot data=outdis; plot y*x=_into_; run;

```

The results of the DISCRIM procedure can be seen from the output and graph that is produced.

#### Diagnosis

##### The DISCRIM Procedure

Observations	100	DF Total	99
Variables	2	DF Within Classes	97
Classes	3	DF Between Classes	2

Class Level Information

dx	Variable Name	Frequency	Weight	Proportion	Prior Probability
Benign	Benign	70	70.0000	0.700000	0.333333
Cancer	Cancer	5	5.0000	0.050000	0.333333
Infection	Infection	25	25.0000	0.250000	0.333333

Diagnosis

The DISCRIM Procedure

Classification Summary for Calibration Data: WORK.PROSTATE  
 Resubstitution Summary using Linear Discriminant Function

Generalized Squared Distance Function

$$D_j(X) = (X - X_j)' \text{COV}_j^{-1} (X - X_j)$$

Posterior Probability of Membership in Each dx

$$\text{Pr}(j|X) = \frac{\exp(-.5 D_j(X))}{\sum_k \exp(-.5 D_k(X))}$$

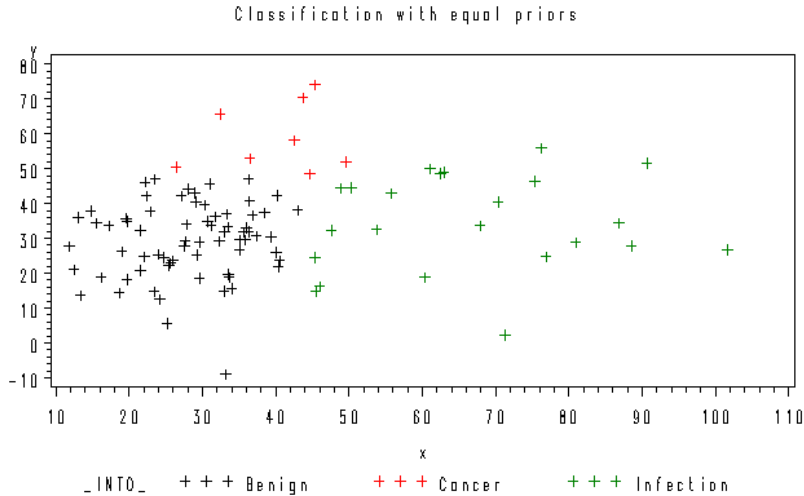
Number of Observations and Percent Classified into dx

From dx	Benign	Cancer	Infection	Total
Benign	65 92.86	3 4.29	2 2.86	70 100.00
Cancer	0 0.00	5 100.00	0 0.00	5 100.00
Infection	4 16.00	0 0.00	21 84.00	25 100.00
Total	69 69.00	8 8.00	23 23.00	100 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for dx

Benign	Cancer	Infection	Total
--------	--------	-----------	-------

Rate	0.0714	0.0000	0.1600	0.0771
Priors	0.3333	0.3333	0.3333	



The following DATA step creates a decision data set containing prior probabilities and a revenue matrix. The revenue matrix indicates the benefit of each treatment. The costs of each treatment (such as bad side effects) will be specified later in a DECISION statement.

The variables are: EQPRIOR = The prior probabilities used by DISCRIM (equal, by default), PRIOR = The known proportions from the population, NOTHING = The benefit of doing nothing (0), ANTIBIOT = The benefit of using antibiotics (cures infection, no benefit for other diseases), and SURGERY = The benefit of surgery (cures all diseases)

```
data rx(type=revenue);
  input dx $10. eqprior prior nothing antibiotic surgery;
  datalines;
Benign          .33    70    0    0    5
Infection       .33    25    0    10   10
Cancer          .33    5    0    0   100
;
```

Use PROC DECIDE to assign a treatment to each patient.

The cost associated with each treatment is: NOTHING = 0 cost, ANTIBIOT = 5 for possible bad side effects, and SURGERY = 20 for possible severe side effects.

```
proc decide data=outdis out=outdec outstat=sumdec;
  target dx;
  posteriors benign infection cancer;
  decision decdata=rx
    oldpriorvar=eqprior priorvar=prior
    decvars=nothing antibiotic surgery
    cost=      0      5      20;
run;

title2 'Treatment: Cost of surgery=20';
proc print data=sumdec label; run;
```

```
proc gplot data=outdec; plot y*x=d_rx; run;
```

The output shows the total and average profits.

```
'Treatment: Cost of surgery=20'
```

	Total Profit for RX	Average Profit for RX
Obs		
1	470	4.7

Some patients may regard the side effects of surgery as more severe than other patients. If we had information from each patient regarding his evaluation of the severity of side effects, we could include a cost variable for surgery (or the other treatments, too) in the data set containing patient scores. However, for illustrative purposes, let's just increase the cost of surgery from 20 to 50 and see what happens across all patients.

```
proc decide data=outdis out=outdec2 outstat=sumdec2;
  target dx;
  posteriors benign infection cancer;
  decision decdata=rx
    oldpriorvar=eqprior priorvar=prior
    decvars=nothing antibiotic surgery
    cost=      0      5      50;
run;

title2 'Treatment: Cost of surgery=50';
proc print data=sumdec2 label; run;
proc gplot data=outdec2; plot y*x=d_rx; run;
```

The new average and total profits are shown in the output, given an increase in the cost of surgery from 20 to 50.

```
'Treatment: Cost of surgery=50'
```

	Total Profit for RX	Average Profit for RX
Obs		
1	285	2.85

---

## References

- Berger, J. O. (1980), *Statistical Decision Theory and Bayesian Analysis*, Second Edition, New York: Springer-Verlag.
- Clemen, R. T. (1991), *Making Hard Decisions: An Introduction to Decision Analysis*, Boston: PWS-Kent.
- DeGroot, M. H. (1970), *Optimal Statistical Decisions*, New York: McGraw-Hill.

- Robert, C. P. (1994), *The Bayesian Choice, a Decision Theoretic Motivation*, New York: Springer-Verlag.
- Savage, L. J. (1972), *The Foundations of Statistics*, Second Revised Edition, New York: Dover.