Nonlife Insurance Risk Classification Using Categorical Embedding

Peng Shi, ACAS, FSA, University of Wisconsin-Madison

Kun Shi, FCAS, Southwest Airline

Motivation

- Risk classification: categorical variables with a large number of levels
- High cardinality can be a challenge in actuarial methods
  - Higher likelihood of sparse data
  - Inherent relation between levels is ignored
  - Unrealistic amount of computational resource

This Talk

- We present the method of categorical embedding

- Discuss several actuarial applications
  - Single insurance risk
  - Dependent insurance risks
  - Pricing new risks

Discussion

- What is the current practice?

- Alternative strategies
  - Credibility theory
  - Generalized linear mixed-effects model
  - Regularized regression

Outline

- Overview of neural network
- Network in GLM
- Categorical embedding
- Applications

Overview of NN

- The concept of ANNS traces back to the 1940s, it is now a powerful machine learning tool

- ANNs are created by combining multiple artificial neurons and represented by Input layer, Hidden layer, and Output layer.

  - Input layer: input variables
  - Hidden layer: learning
  - Output layer: predicted value

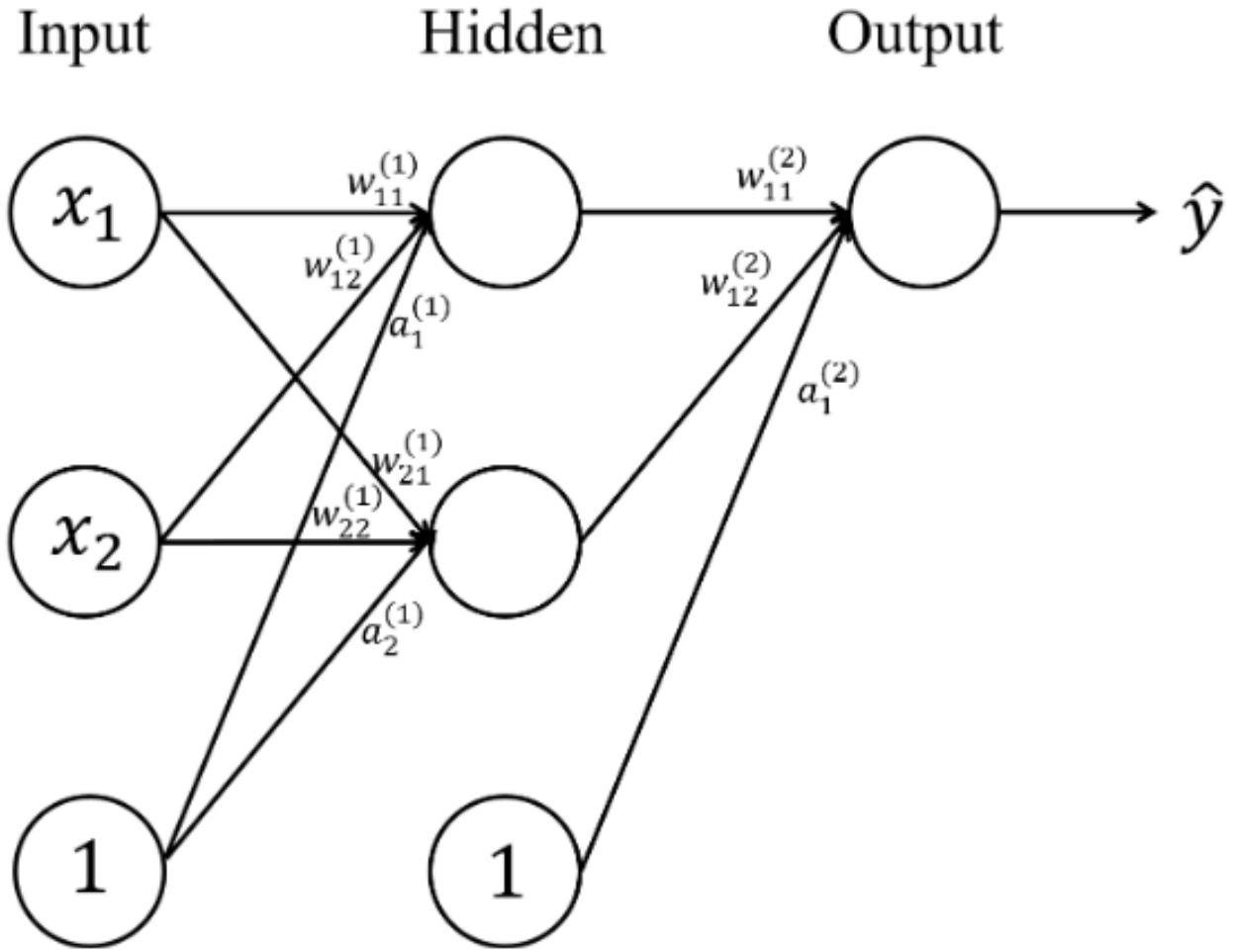
Overview of NN

- Bias neuron
- Weights
- Net input

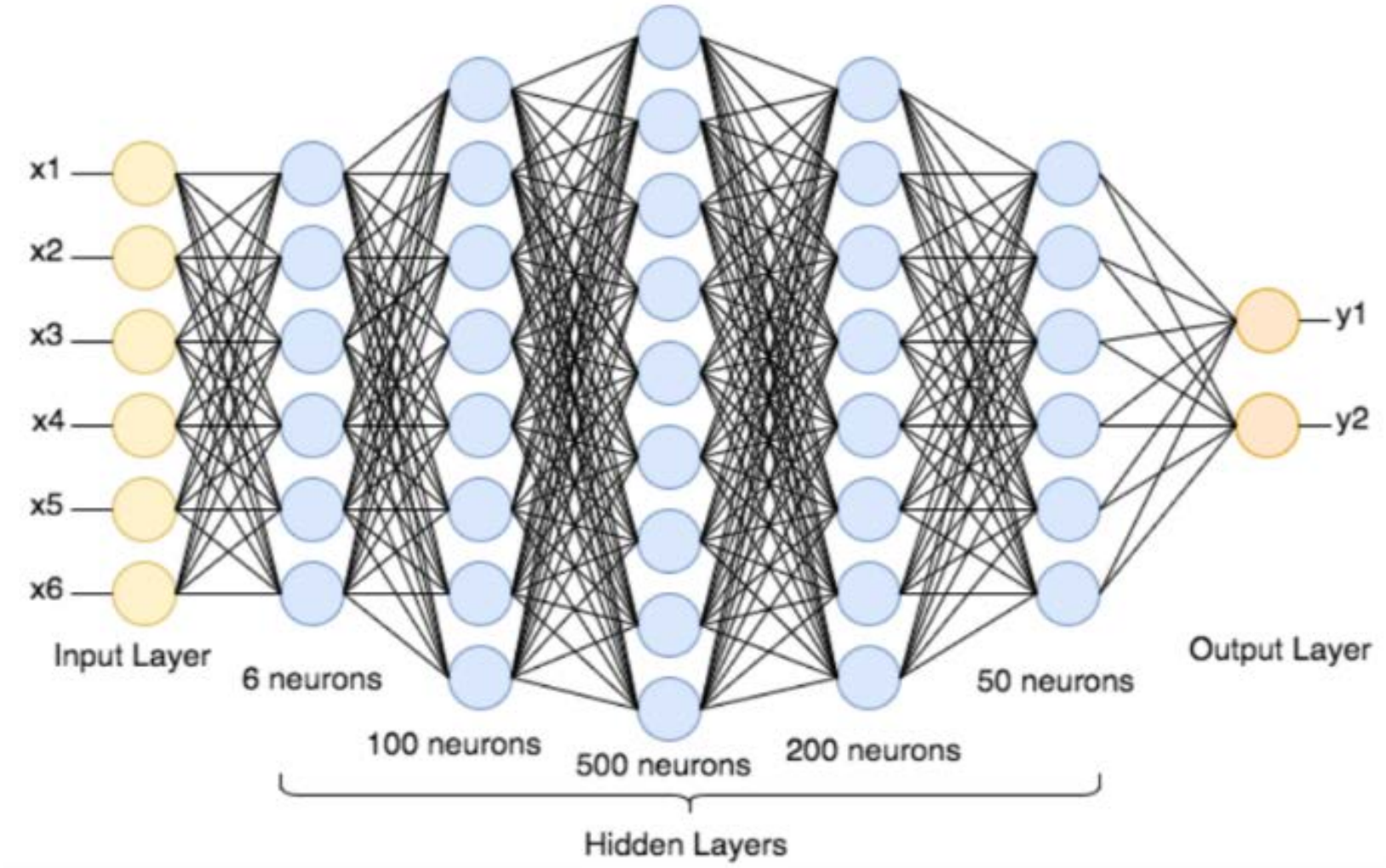
=bias + $\sum(\text{weight} \times \text{input})$

- Activation function
- Output (activations)

= $h(\text{net input})$



Deep Network



Overview of NN

For the jth neuron in layer l:

- Net input = bias + $\sum$ (weight $\times$ input)

$$u_j^{(l)} = a_j^{(l)} + \sum_{k=1}^{N_{l-1}} z_k^{(l-1)} w_{jk}^{(l)}$$

where $a_j^{(l)}$ is a bias term and w_{jk} is the weight.

- Output, known as activations:

$$z_j^{(l)} = h^{(l)}(u_j^{(l)})$$

where $h^{(l)}(\cdot)$ is called activation function.

Activation Function

Examples of Nonlinear activation functions:

- Sigmoid/Logistic: $h(u) = \dfrac{1}{1+e^{-u}}$
- TanH/Hyperbolic Tangent: $h(u) = \dfrac{e^{u}-e^{-u}}{e^{u}+e^{-u}}$
- ReLU (Rectified Linear Unit) $h(u) = \max\{0, u\} = \begin{cases} u & \text{if } u > 0 \\ 0 & \text{if } u \leq 0 \end{cases}$
- Softsign: $h(u) = \dfrac{z}{1+|u|}$

NN in GLM

- Recall that in GLM

$$y_i|\boldsymbol{x}_i \sim \text{Exp}\ (\mu_i, \phi)$$

$$E(y_i|\boldsymbol{x}_i)=\mu_i,\ Var(y_i|\boldsymbol{x}_i)=\phi\ \mu_i V(\mu_i)$$

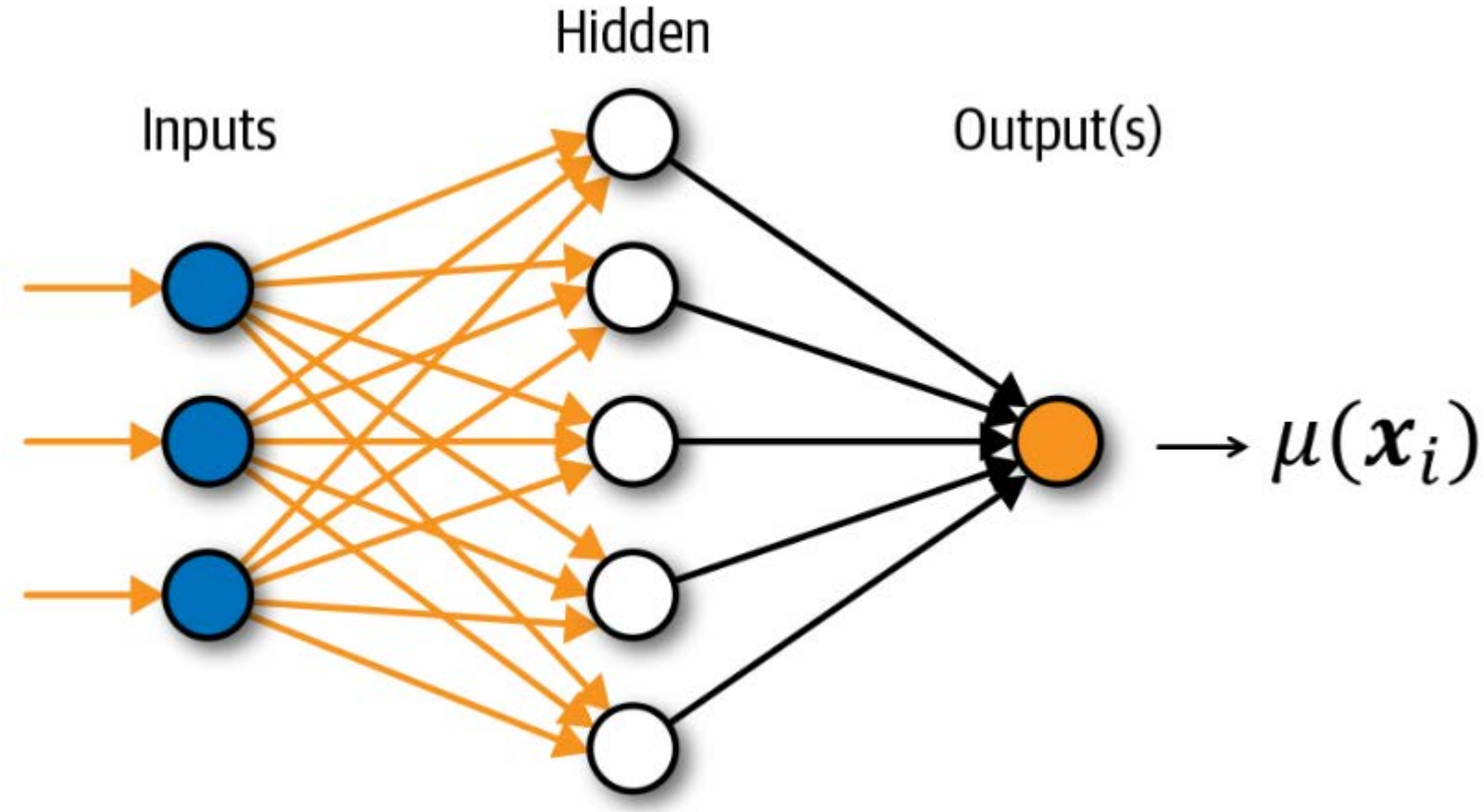
$$\text{where } \mu_i = \mu(\boldsymbol{x}_i)$$

- The choice of the variance function drives many inference properties, not the choice of the distribution.

Distribution	Variance Function $V(\mu)$
Normal	1
Bernoulli	$\mu(1-\mu)$
Poisson	μ
Gamma	μ^2
Tweedie	$\mu^p\ (p \in (1,2))$
Inverse Gaussian	μ^3

NN in GLM

- We use the output from the feedforward network to model $\mu(\boldsymbol{x}_i)$



NN in GLM

- We use the output from the feedforward network to model $\mu(\boldsymbol{x}_i)$

$$\mu(\boldsymbol{x}_i) = z_{i1}^{(L)} = h^{(L)}(u_{i1}^{(L)}) = h^{(L)}(a_1^{(L)} + \sum_{k=1}^{N_{L-1}} z_{ik}^{(L-1)} w_{1k}^{(L)})$$

$$= g^{-1}(\beta_0 + \beta_1 \tilde{x}_{i1} + \cdots \beta_q \tilde{x}_{iq})$$

- $\tilde{x}_{i1}, \ldots, \tilde{x}_{iq}$ are engineered features

$$\tilde{x}_{ik} = z_{ik}^{(L-1)} = h^{(L-1)}(h^{(L-2)}(\cdots h^{(1)}(a_j^{(1)} + \sum_{k=1}^{p} x_k w_{jk}^{(1)})))$$

Categorical Embedding

- One-hot encoding

One-hot encoding can be formulated as a function h that maps the categorical variable into a binary vector of length K:

$$h : x \mapsto \boldsymbol{\delta} = (\delta_{x,c_1}, \ldots, \delta_{x,c_K})',$$

where δ_{x,c_k}, for $k = 1, \ldots, K$, is the Kronecker delta which equals 1 if $x = c_k$ and 0 otherwise.

Categorical Embedding

- One-hot encoding

One-hot encoding can be formulated as a function h that maps the categorical variable into a binary vector of length K:

$$h : x \mapsto \boldsymbol{\delta} = (\delta_{x,c_1}, \ldots, \delta_{x,c_K})',$$

where δ_{x,c_k}, for $k = 1, \ldots, K$, is the Kronecker delta which equals 1 if $x = c_k$ and 0 otherwise.

- Issues with neural networks
  - Continuity concern
  - High cardinality

Categorical Embedding

- The method maps each categorical variable into a real-valued representation in the Euclidean space.
  - In the embedding space, the categories with similar effects are close to each other

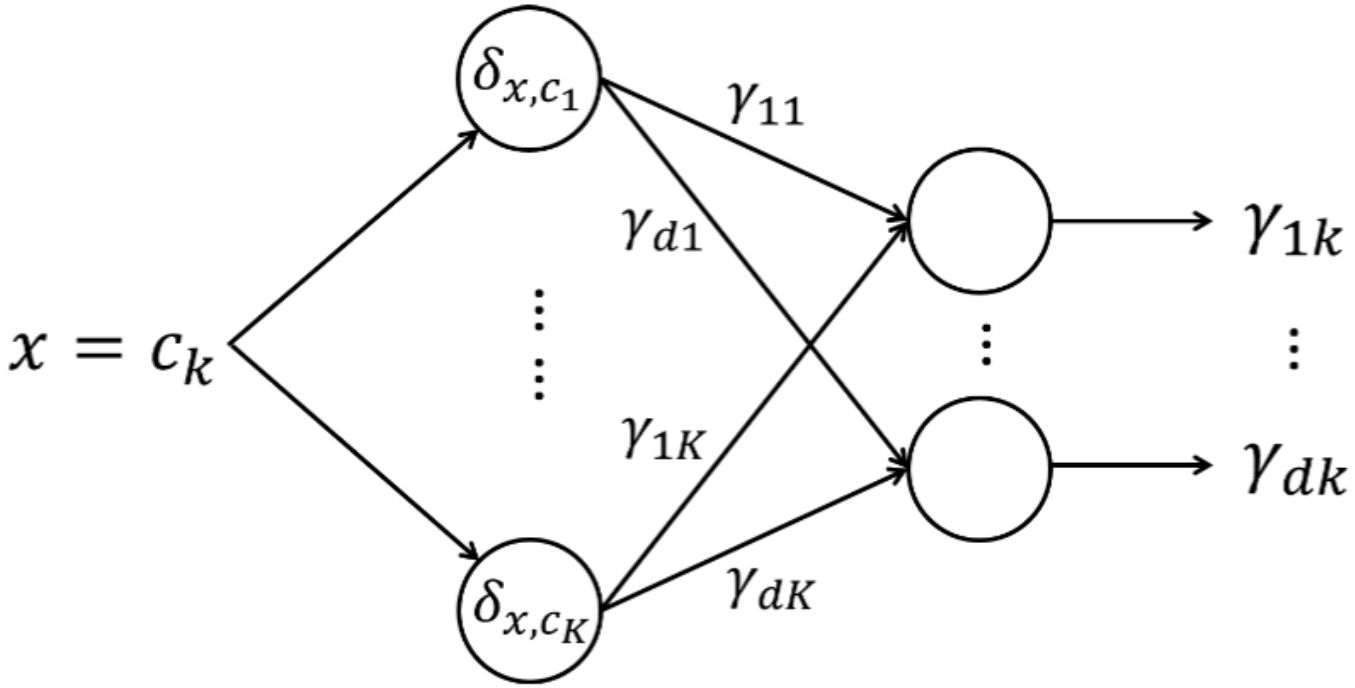
Be more specific, for data point with $x_i = c_k$, we note:

$$e(x_i) = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1K} \\ \vdots & \ddots & \vdots \\ \gamma_{d1} & \cdots & \gamma_{dK} \end{pmatrix} \times \begin{pmatrix} \delta_{x_i,c_1} \\ \vdots \\ \delta_{x_i,c_K} \end{pmatrix} = \begin{pmatrix} \gamma_{1k} \\ \vdots \\ \gamma_{dk} \end{pmatrix}.$$

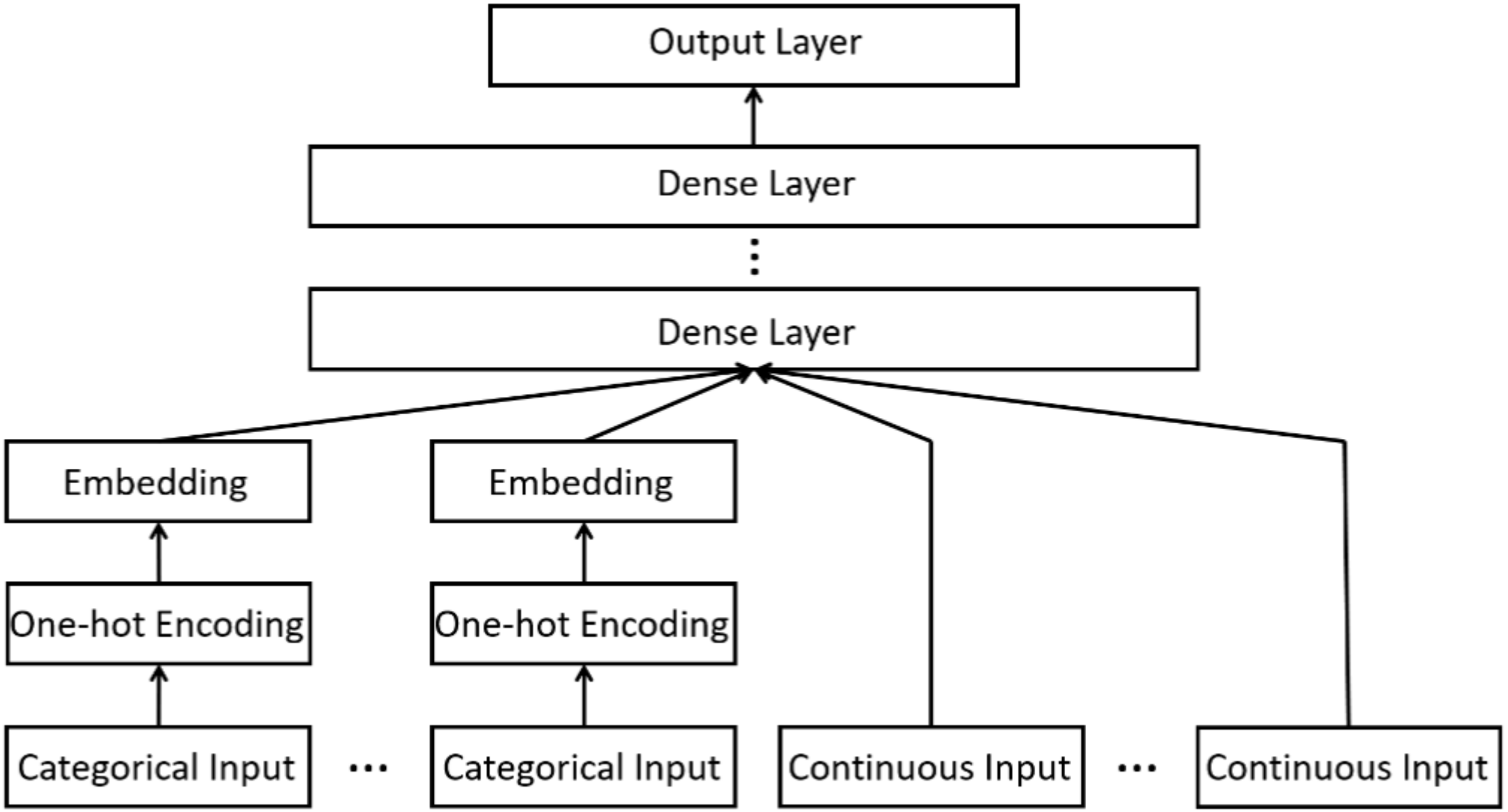
The dimension of embedding space is bounded between 1 and $K-1$, i.e. $1 \leq d \leq K-1$.

Categorical Embedding

- The embeddings can be automatically learned by a neural network in the supervised training process.
  - Add an embedding layer, an extra layer between the input layer and the hidden layer
  - Treat the embedding matrix as the weight parameters of the embedding neurons



Categorical Embedding



Actuarial Applications

- In predictive models, it is one way to incorporate categorical variables
  - Dimension reduction
  - Variable selection

- Learned embeddings could be the interest.

Data

- The insurance claims dataset is obtained from the local government property insurance fund of Wisconsin
  - We examine the building and contents insurance that covers damage to both physical structures and items inside
  - There are over one thousand entities observed during years 2006-2013, resulting in 8,880 policy-year observations.

Variable	Description
EntityType	Entity type of the policyholder (city, county, school, town, or village, or other)
County	County code of the geographical region of the property (72 categories)
Coverage	Amount of insurance coverage
Deductible	Amount of deductible

Data

- We consider a binary outcome that measures the frequency of insurance claims by peril

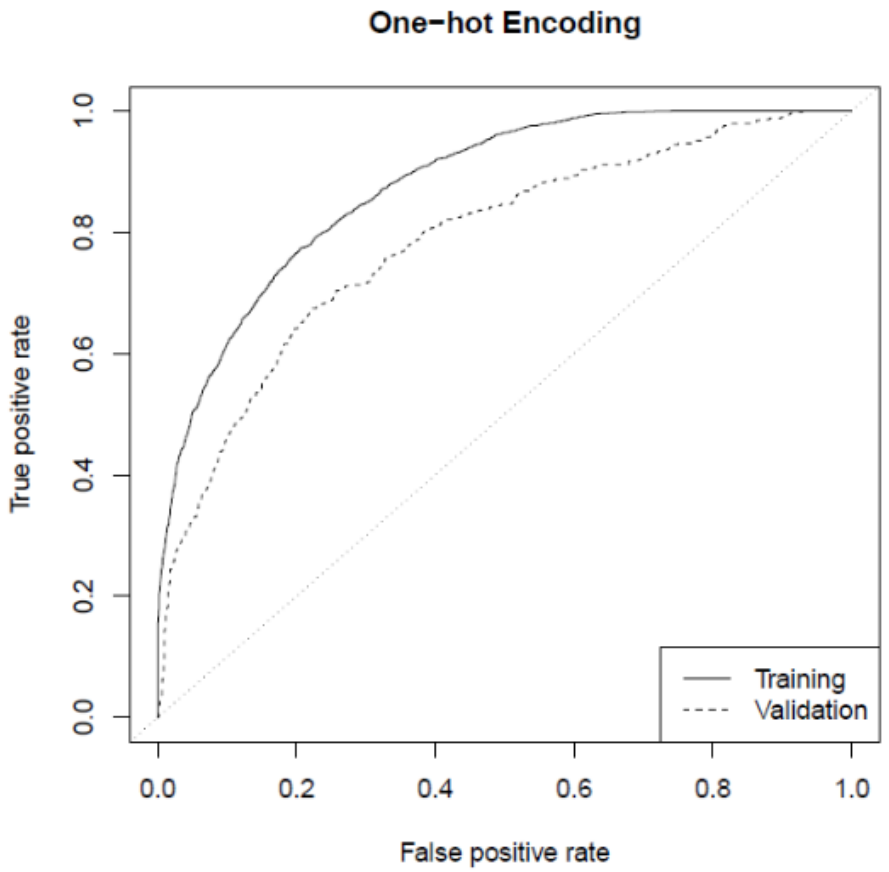
	Overall	Entity Type					
		City	County	School	Town	Village	Misc
Claim Probability	0.291	0.522	0.743	0.302	0.077	0.263	0.116
Peril Type							
Fire	0.158	0.333	0.457	0.133	0.044	0.138	0.057
Water	0.124	0.225	0.464	0.114	0.020	0.103	0.039
Other	0.116	0.206	0.362	0.145	0.023	0.068	0.037
Number of Obs	8,880	1,247	541	2,469	1,510	2,117	996

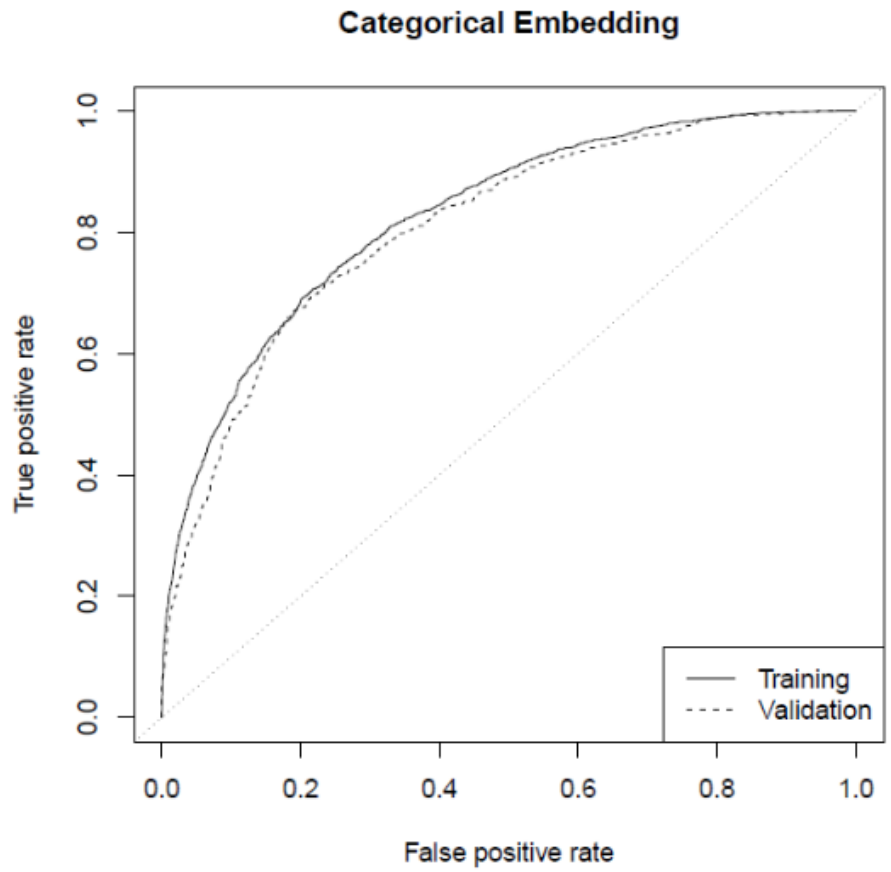
Rating Classes for A Single Risk

- Consider a single risk:
  - Treat the open-peril property insurance as an umbrella policy
  - Define the claim frequency as a risk measurement for the aggregate claims from all perils

- Train-test split to assess the performance

Rating Classes for A Single Risk





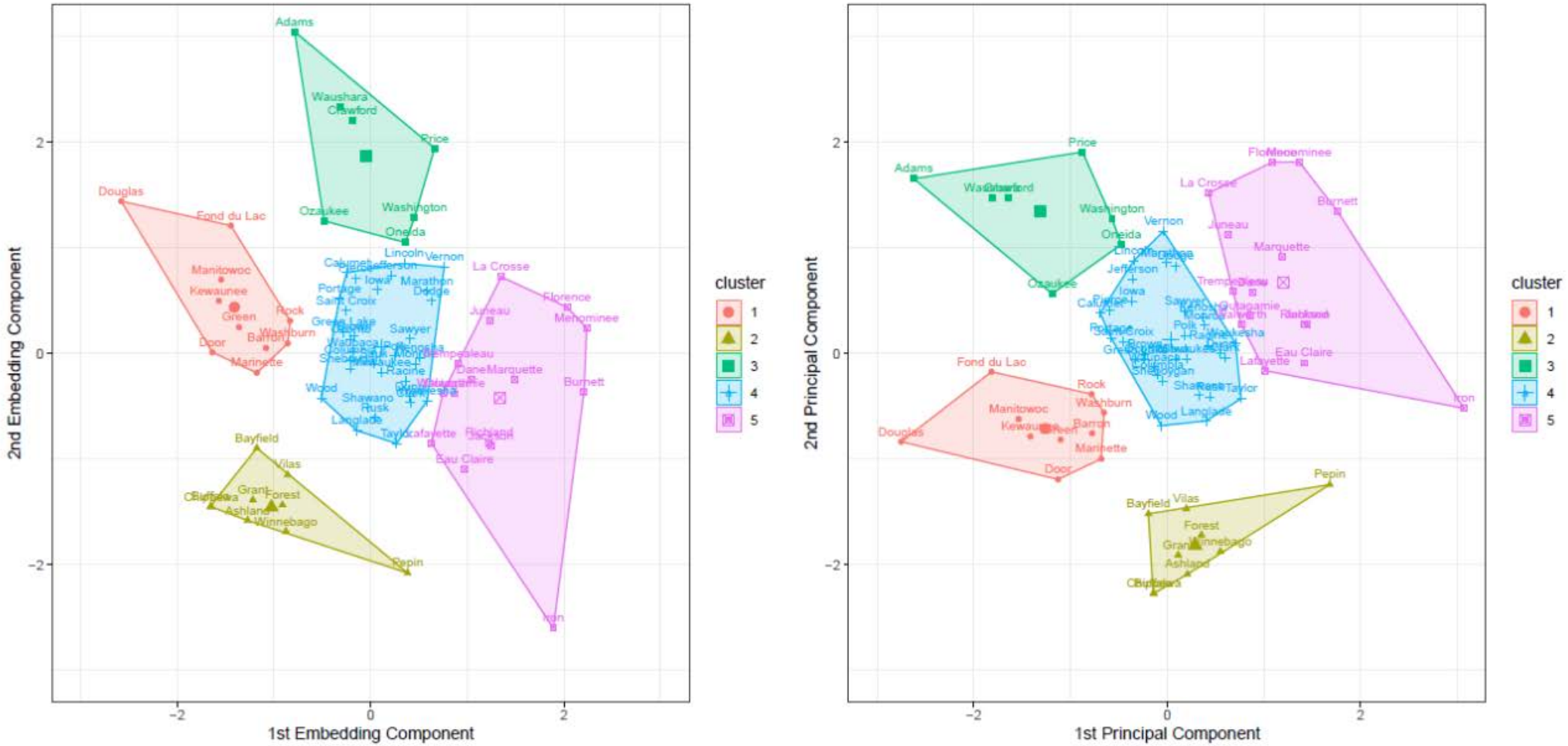
Rating Classes for A Single Risk

- Comparison between one-hot encoding and categorical embedding

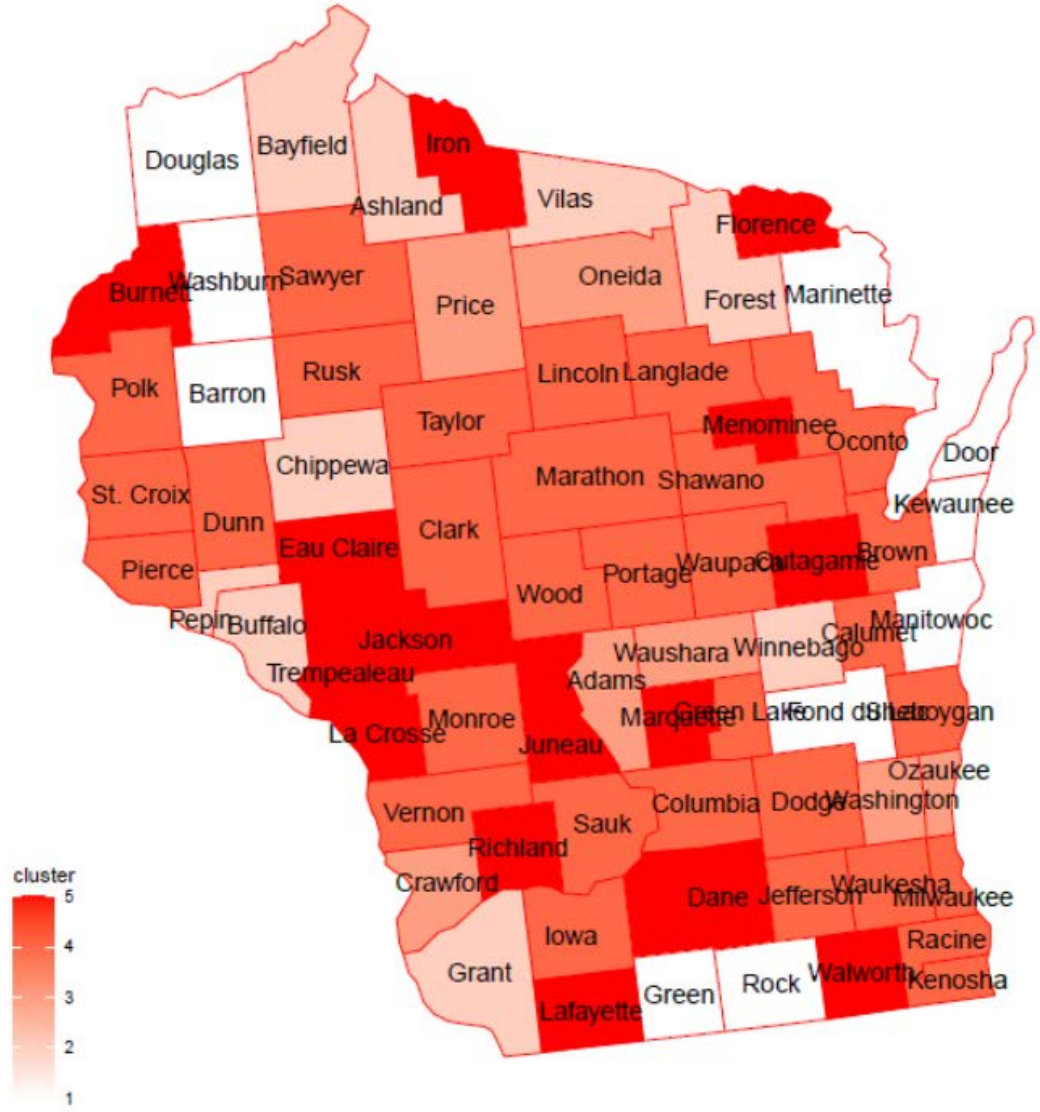
	One-hot Encoding	Categorical Embedding
Simple Gini	41.534 (1.745)	45.204 (1.619)
One-hot Encoding		21.152 (2.050)
Categorical Embedding	3.049 (2.004)	

† Standard errors are reported in parentheses.

Rating Classes for A Single Risk



Rating Classes for A Single Risk



Dependent Risks

- We consider a multivariate risk context
  - Each peril is viewed as a single risk
  - Let Z_j be the outcome for jth peril. Out interest is $Y = (Z_1, Z_2, Z_3)$
- We are interested in quantity:

$$\Pr(Z_1, Z_2, Z_3) \neq \Pr(Z_1)\Pr(Z_2)\Pr(Z_2)$$

- We transform the modeling of Y to a multi-class classification, and consider a multi-output network for Y

$$\hat{y}_j = g^{(L)}\left(\boldsymbol{u}^{(L)}\right) = \frac{\exp\left\{u_j^{(L)}\right\}}{\sum_{j=1}^{8}\exp\left\{u_j^{(L)}\right\}}, \quad \text{for} \quad j = 1, \ldots, 8$$

Dependent Risks

- We use dependence ratio to describe the relationship among perils

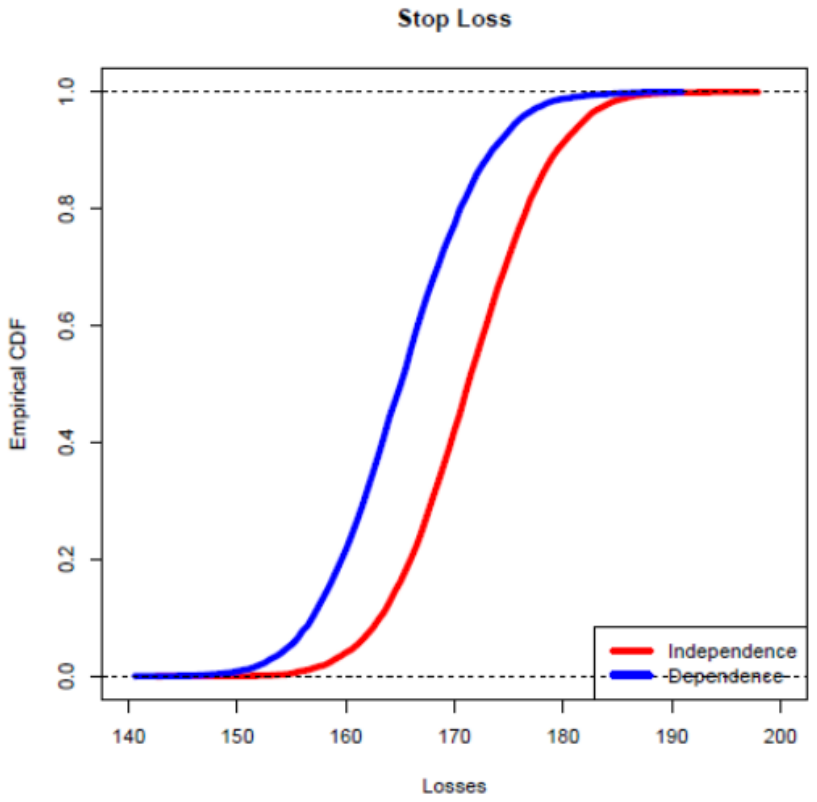
$$\rho(z_1, z_2, z_3) = \frac{\Pr(Z_1 = z_1, Z_2 = z_2, Z_3 = z_3)}{\Pr(Z_1 = z_1)\Pr(Z_2 = z_2)\Pr(Z_3 = z_3)}$$

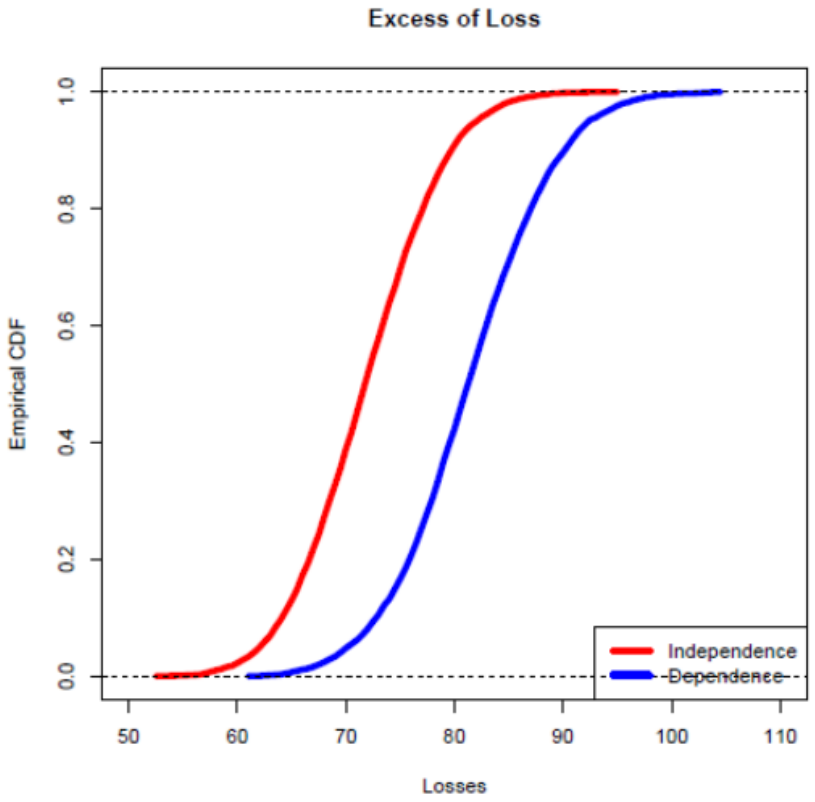
	Observed Value	Association Ratio	Fitted Value	
			Independent	Dependent
$(0,0,0)$	4,745	1.450	4,687	4,757
$(0,0,1)$	410	0.830	431	401
$(0,1,0)$	444	0.849	478	453
$(1,0,0)$	557	0.963	599	550
$(0,1,1)$	102	1.264	116	101
$(1,0,1)$	157	1.810	143	152
$(1,1,0)$	178	1.862	176	179
$(1,1,1)$	165	11.597	128	164
$\chi^2 - statistic$			20.883	0.686

Dependent Risks

- The insurer's retained loss under 1) stop loss; 2) Excess of loss

$$\text{Stop loss}: R_1 = \min\{S, d_1\}$$
$$\text{Excess of loss}: R_2 = \max\{S - d_2, 0\}$$





Pricing New Risks

- Suppose that the insurer has only provided coverage for water and other perils during years 2006-2011. Starting from year 2012, the insurer plans to offer fire coverage as well.
- We demonstrate the idea of transfer learning using the categorical variable county.
  - *Learn the embeddings from single peril: water or other*
  - *Learn the embeedings from the joint bi-peril model: water and other*
- Two comparisons
  - One-hot encoding for ``county''
  - Arbitrary grouping for ``county''

Pricing New Risks

- We use data in years 2012-2013 to do back-testing
  - *Fitted model using data in 2012*

	Based on Embedding Cluster				Based on Embedding Matrix		
	Estimate	Std. Error	p-value		Estimate	Std. Error	p-value
Learned from Water Peril							
(Intercept)	-2.557	0.855	0.003	(Intercept)	-1.439	0.797	0.071
City	1.327	0.525	0.012	City	1.281	0.525	0.015
County	1.495	0.565	0.008	County	1.357	0.563	0.016
School	-0.006	0.527	0.991	School	-0.046	0.526	0.930
Town	1.328	0.654	0.042	Town	1.357	0.653	0.038
Village	1.345	0.530	0.011	Village	1.330	0.528	0.012
Coverage	0.889	0.109	0.000	Coverage	0.926	0.111	0.000
Deductible	-0.485	0.095	0.000	Deductible	-0.482	0.094	0.000
Region2	0.982	0.414	0.018	Embedding1	1.430	0.744	0.055
Region3	1.120	0.399	0.005	Embedding2	-0.389	0.283	0.169
χ^2-statistic		6.029		χ^2-statistic		6.150	
Loglik		-371.079		Loglik		-372.769	

Pricing New Risks

- We use data in years 2012-2013 to do back-testing
  - *Test using data in 2013*

	Embedding Cluster			Embedding Matrix		
	Water	Other	Water+Other	Water	Other	Water+Other
Average	50.973	51.502	51.280	51.068	52.162	51.725
	(2.906)	(2.887)	(2.878)	(2.887)	(2.829)	(2.851)
County	20.892	21.393	22.640	20.504	23.622	22.496
	(4.029)	(3.981)	(3.931)	(4.044)	(3.869)	(3.904)

† Standard errors are reported in parentheses.

Pricing New Risks

- Consider case where the counties are arbitrarily grouped
- We compare three methods in terms
  - 1) directly use embedding vectors as predictors
  - 2) 3 groups of counties from clustering the embeddings
  - 3) randomly assign counties into 3 groups, replicate 500 times
- Embeddings are learned from the joint model for water and other perils
  - In-sample (AIC): 1) and 2) outperforms 3) 94% and 97% times respectively
  - Hold-out (AUC): 1) and 2) outperforms 3) 95% and 63% times respectively

Summary

- Introduced the method of categorical embedding
- Discussed several actuarial applications
- Two distinctive aspects of the method
  - *The method is viewed as a way to incorporate categorical input variables in deep neural networks*
  - *The neural network is viewed as a vehicle for computing the embeddings for categorical input variables*